



(11)

EP 3 611 725 A1

(12)

EUROPEAN PATENT APPLICATION
published in accordance with Art. 153(4) EPC

(43) Date of publication:
19.02.2020 Bulletin 2020/08

(51) Int Cl.:
G10L 17/04 (2013.01) G10L 15/16 (2006.01)

(21) Application number: **18880575.8**

(86) International application number:
PCT/CN2018/115704

(22) Date of filing: **15.11.2018**

(87) International publication number:
WO 2019/100998 (31.05.2019 Gazette 2019/22)

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR
Designated Extension States:
BA ME
Designated Validation States:
KH MA MD TN

- **YU, Meng**
Shenzhen
Guangdong 518057 (CN)
- **LUO, Min**
Shenzhen
Guangdong 518057 (CN)
- **SU, Dan**
Shenzhen
Guangdong 518057 (CN)

(30) Priority: **24.11.2017 CN 201711191604**

(71) Applicant: **TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED**
Shenzhen, Guangdong 518057 (CN)

(74) Representative: **Gunzelmann, Rainer**
Wuesthoff & Wuesthoff
Patentanwlte PartG mbB
Schweigerstrae 2
81541 Mnchen (DE)

(72) Inventors:
• **CHEN, Lianwu**
Shenzhen
Guangdong 518057 (CN)

(54) **VOICE SIGNAL PROCESSING MODEL TRAINING METHOD, ELECTRONIC DEVICE, AND STORAGE MEDIUM**

(57) Embodiments of the present invention provide a speech signal processing model training method, an electronic device and a storage medium. The embodiments of the present invention determines a target training loss function based on a training loss function of each speech signal processing task; inputs a task input feature of each speech signal processing task into a to-be-trained multi-task neural network, and updates parameters of a shared layer and each task layer of the to-be-trained multi-task neural network by minimizing the target training loss function as a training objective, until the to-be-trained multi-task neural network converges, to obtain a speech signal processing model. Through the multi-task neural network, the computational complexity is reduced, and the efficiency of training the speech signal processing model is improved.

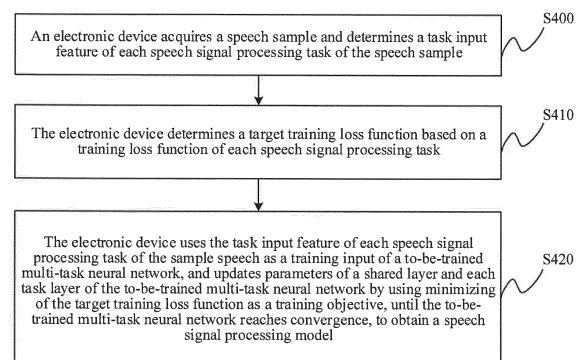


FIG. 10

EP 3 611 725 A1

Description

RELATED APPLICATION

[0001] This application claims priority to Chinese Patent Application No. 201711191604.9, entitled "SPEECH SIGNAL PROCESSING MODEL TRAINING METHOD, APPARATUS, ELECTRONIC DEVICE AND STORAGE MEDIUM" filed with the China National Intellectual Property Administration on November 24, 2017, which is incorporated by reference in its entirety.

FIELD OF THE TECHNOLOGY

[0002] Embodiments of the present invention relate to the field of speech processing technologies, and in particular, to a speech signal processing model training method, an electronic device and a storage medium for processing speech signal.

BACKGROUND OF THE DISCLOSURE

[0003] With the development of speech recognition technologies, the performance of the speech signal processing technology of terminals is particularly important for maintaining a high speech recognition rate in complex environments. Currently, a common speech recognition process is as follows: a terminal performs speech signal processing on an input multi-channel speech, outputs a single-channel speech, and then sends the single-channel speech to a speech backend server for speech recognition.

[0004] A conventional speech signal processing procedure generally includes a plurality of speech signal processing tasks. The plurality of speech signal processing tasks collaboratively processes an input multi-channel speech in a progressive manner, and outputs a single-channel speech. Using an intelligent speaker scenario as an example, FIG. 1 shows a conventional speech signal processing procedure in a terminal. The process includes a plurality of speech signal processing tasks. The plurality of speech signal processing tasks may specifically include: an echo cancellation task, a speech detection task, a speech direction detection task, a microphone array enhancement task, a single-channel noise reduction task, a de-reverberation task and the like. After an input multi-channel speech is collaboratively processed by the plurality of speech signal processing tasks, a single-channel speech may be output, thus completing the speech signal processing by the terminal.

[0005] Deep learning technologies, such as neural network technologies are being applied to more technical fields. To improve the speech signal processing performance of the terminal, a technology of optimizing the speech signal processing procedure of the terminal by using a neural network has been adopted in the art. The technology is to train a speech signal processing model by using the neural network and use the speech signal

processing model to replace or assist the conventional speech signal processing procedure in the terminal, and to improve the speech signal processing performance of the terminal. Therefore, the training of the speech signal processing model based on the neural network is of important technical significance in improving the speech signal processing performance.

[0006] Currently, the training of the speech signal processing model based on the neural network is confronted with the problem that the training efficiency of the speech signal processing model is low due to a large quantity of speech signal processing tasks involved in the speech signal processing procedure and complex computations involved in the training.

SUMMARY

[0007] In view of this, embodiments of the present invention provide a speech signal processing model training method, an electronic device and a storage medium, to reduce the computational complexity of training the speech signal processing model, and improve the efficiency of training the speech signal processing model.

[0008] In order to achieve the foregoing objective, the embodiments of the present invention provide the following technical solutions:

[0009] According to an aspect, an embodiment of the present invention provides a speech signal processing model training method, applied to an electronic device, and including:

acquiring a sample speech and determining a task input feature of each speech signal processing task of the sample speech;

determining a target training loss function based on a training loss function of each speech signal processing task; and

using the task input feature of each speech signal processing task of the sample speech as a training input of a to-be-trained multi-task neural network, and updating parameters of a shared layer and each task layer of the to-be-trained multi-task neural network by minimizing the target training loss function as a training objective, until the to-be-trained multi-task neural network converges, to obtain a speech signal processing model,

the to-be-trained multi-task neural network including: the shared layer and the task layer corresponding to each speech signal processing task.

[0010] According to another aspect, an embodiment of the present invention further provides a speech signal processing model training apparatus, applied to an electronic device, and including:

a task input feature determining module, configured to acquire a sample speech and determine a task input feature of each speech signal processing task of the sample speech;

a target loss function determining module, configured to determine a target training loss function based on a training loss function of each speech signal processing task; and

a model training module, configured to use the task input feature of each speech signal processing task of the sample speech as a training input of a to-be-trained multi-task neural network, and update parameters of a shared layer and each task layer of the to-be-trained multi-task neural network by minimizing the target training loss function as a training objective, until the to-be-trained multi-task neural network converges, to obtain a speech signal processing model,

the to-be-trained multi-task neural network including: the shared layer and the task layer corresponding to each speech signal processing task.

[0011] According to another aspect, an embodiment of the present invention further provides an electronic device, including: at least one memory and at least one processor; the memory storing a program, the processor invoking the program stored by the memory, and the program being configured for:

acquiring a sample speech and determining a task input feature of each speech signal processing task of the sample speech;

determining a target training loss function based on a training loss function of each speech signal processing task; and

using the task input feature of each speech signal processing task of the sample speech as a training input of a to-be-trained multi-task neural network, and updating parameters of a shared layer and each task layer of the to-be-trained multi-task neural network by minimizing the target training loss function as a training objective, until the to-be-trained multi-task neural network converges, to obtain a speech signal processing model,

the to-be-trained multi-task neural network including: the shared layer and the task layer corresponding to each speech signal processing task.

[0012] According to another aspect, an embodiment of the present invention further provides a storage medium, storing a program configured to be executed by a processor, the program being used for:

acquiring a sample speech and determining a task input feature of each speech signal processing task of the sample speech;

determining a target training loss function based on a training loss function of each speech signal processing task; and

using the task input feature of each speech signal processing task of the sample speech as a training input of a to-be-trained multi-task neural network, and updating parameters of a shared layer and each task layer of the to-be-trained multi-task neural network by minimizing the target training loss function as a training objective, until the to-be-trained multi-task neural network converges, to obtain a speech signal processing model,

the to-be-trained multi-task neural network including: the shared layer and the task layer corresponding to each speech signal processing task.

[0013] The embodiments of the present invention determine the target training loss function based on training loss functions of a plurality of speech signal processing tasks, uses the task input features of the plurality of speech signal processing tasks as the training input of the multi-task neural network, and trains the to-be-trained multi-task neural network by minimizing the target training loss function as a training objective, to obtain the speech signal processing model. The multi-task neural network includes the shared layer and the task layer corresponding to each speech signal processing task, and the speech signal processing model is obtained through training the multi-task neural network, instead of training the neural network respectively for each speech signal processing task, thereby effectively reducing the computational complexity of training the speech signal processing model, and improving the efficiency of training.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] To describe the technical solutions in the embodiments of the present invention or in the related technology more clearly, the following briefly describes the accompanying drawings required for describing the embodiments or the related technology. Apparently, the accompanying drawings in the following description show merely embodiments of the present invention, and a person of ordinary skill in the art may still derive other drawings from these accompanying drawings without creative efforts.

FIG. 1 is a schematic diagram of a speech signal processing procedure in the related technology.

FIG. 2 is a schematic diagram of training a speech signal processing model by using a neural network

in the related technology.

FIG. 3 is a schematic structural diagram of a multi-task neural network according to an embodiment of the present invention.

FIG. 4 is another schematic structural diagram of a multi-task neural network according to an embodiment of the present invention.

FIG. 5 is a flowchart of a speech signal processing model training method according to an embodiment of the present invention.

FIG. 6 is a schematic diagram of training a speech signal processing model.

FIG. 7 is another flowchart of a speech signal processing model training method according to an embodiment of the present invention.

FIG. 8 is another schematic diagram of training a speech signal processing model.

FIG. 9 is yet another flowchart of a speech signal processing model training method according to an embodiment of the present invention.

FIG. 10 is still another flowchart of a speech signal processing model training method according to an embodiment of the present invention.

FIG. 11 is a schematic diagram of an application scenario of a speech signal processing model.

FIG. 12 is a schematic diagram of using an output result of a speech signal processing model.

FIG. 13 is a structural block diagram of a speech signal processing model training apparatus according to an embodiment of the present invention.

FIG. 14 is another structural block diagram of a speech signal processing model training apparatus according to an embodiment of the present invention.

FIG. 15 is yet another structural block diagram of a speech signal processing model training apparatus according to an embodiment of the present invention.

FIG. 16 is a block diagram of a hardware structure of an electronic device.

DESCRIPTION OF EMBODIMENTS

[0015] FIG. 2 is a schematic diagram of training a

speech signal processing model by using a neural network in the related technology. As shown in FIG. 2, a neural network is constructed respectively for each speech signal processing task in a speech signal processing procedure. Each neural network corresponds to a speech signal processing task, and the neural network of each speech signal processing task is trained respectively. When a neural network satisfies a training convergence condition of the corresponding speech signal processing task, the training of the neural network is complete. After the training of each neural network is complete, the trained neural networks are combined to form a speech signal processing model. This process has a problem that the training of a neural network is required for each speech signal processing task and the large quantity of speech signal processing tasks leads to high computational complexity of the training. In addition, the neural networks are independent of each other and no association exists between the speech signal processing tasks, resulting in that the performance of the trained speech signal processing model has limitations.

[0016] Based on this, the embodiments of the present invention are directed to improving a neural network structure of a speech signal processing model and training the speech signal processing model based on the improved neural network structure, thereby reducing the computational complexity of training the speech signal processing model and improving the training efficiency. In addition, an association between the speech signal processing tasks is reflected in the training process, to ensure that the trained speech signal processing model has reliable performance.

[0017] The following clearly and completely describes the technical solutions in the embodiments of the present invention with reference to the accompanying drawings in the embodiments of the present invention. Obviously, the described embodiments are only some embodiments instead of all embodiments of the present invention. All other embodiments obtained by a person of ordinary skill in the art based on the embodiments of the present invention without creative effects shall fall within the protection scope of the present disclosure.

[0018] An embodiment of the present invention provides a novel multi-task neural network. By improving a neural network structure of a speech signal processing model, the multi-task neural network can reduce the computational complexity of training the speech signal processing model and further ensure reliable performance of the speech signal processing model. As shown in FIG. 3, the multi-task neural network may include: a shared layer and a task layer corresponding to each speech signal processing task.

[0019] Optionally, in this embodiment of the present invention, an input of the shared layer may be imported into each task layer, and each task layer outputs a task processing result of the speech signal processing task corresponding to the task layer. The shared layer may reflect an association between the speech signal

processing tasks having common features, and each task layer may reflect a task feature of the corresponding speech signal processing task, so that an output result of each task layer can better reflect a task requirement of the corresponding speech signal processing task.

[0020] Optionally, in this embodiment of the present invention, the shared layer may be defined as a long short term memory (LSTM) network. In an optional example, the shared layer may be a two-layer LSTM network. The task layer may be defined as a fully connected multi layer perceptron (MLP) network, that is, each task layer may be a fully connected MLP network. In an optional example, each task layer may be a single-layer fully connected MLP network.

[0021] Using the plurality of speech signal processing tasks shown in FIG. 1 as an example, the multi-task neural network according to this embodiment of the present invention may include, as shown in FIG. 4:

a shared layer, an echo cancellation task layer, a speech detection task layer, ..., a single-channel noise reduction task layer, and a de-reverberation task layer.

[0022] Certainly, in a specific speech signal processing procedure, the plurality of speech signal processing tasks is not limited to those shown in FIG. 1, and some speech signal processing tasks may be deleted from and/or added to the plurality of speech signal processing tasks shown in FIG. 1. This is not specifically limited in this embodiment of the present invention.

[0023] Based on the multi-task neural network according to this embodiment of the present invention, this embodiment of the present invention can train the multi-task neural network, to obtain a speech signal processing model.

[0024] In an optional implementation of training the speech signal processing model, this embodiment of the present invention can train the multi-task neural network based on all the speech signal processing tasks simultaneously to update parameters of the shared layer and each task layer of the multi-task neural network.

[0025] Optionally, FIG. 5 shows an optional procedure of a speech signal processing model training method according to an embodiment of the present invention. The method may be applied to an electronic device having a data processing capability. The electronic device may be a terminal device having a data processing capability, such as a notebook computer, a personal computer (PC) and the like, or may be a server on a network side. This is not specifically limited in this embodiment of the present invention. Referring to FIG. 5, the procedure may include the following steps.

[0026] Step S100: The electronic device acquires a sample speech and determines a task input feature of each speech signal processing task of the sample speech.

[0027] Optionally, the sample speech may be considered as a sample used to train a speech signal processing model, and the sample speech may be a multi-channel speech. This embodiment of the present invention may

acquire a plurality of sample speeches, and determine, for each sample speech, a task input feature of each speech signal processing task.

[0028] For a plurality of speech signal processing tasks involved in a speech signal processing procedure of a terminal, this embodiment of the present invention may acquire a task input feature of each speech signal processing task respectively for a sample speech. Optionally, the plurality of speech signal processing tasks involved in the speech signal processing procedure of the terminal may be as shown in FIG. 1. Certainly, some speech signal processing tasks may be deleted from and/or other forms of speech signal processing tasks may be added to the plurality of speech signal processing tasks shown in FIG. 1.

[0029] Optionally, for ease of understanding, in an optional example, the plurality of speech signal processing tasks includes an echo cancellation task and a speech detection task. The echo cancellation task may be used to estimate a single-channel speech spectrum, and the speech detection task may be used to estimate a speech presence probability. Therefore, this embodiment of the present invention may acquire a task input feature of the echo cancellation task of the sample speech, where the task input feature is, for example, spectral energy of a noisy single-channel speech of the sample speech and spectral energy of a speech marked as clean. In addition, this embodiment of the present invention may acquire a task input feature of the speech detection task of the sample speech, where the task input feature is, for example, a tag value indicating whether a speech exists in the sample speech. The tag value may be 0 or 1, where 0 indicates the absence of a speech and 1 indicates the presence of a speech.

[0030] Certainly, the speech signal processing tasks described above are examples only, and in practice, the speech signal processing procedure may include more speech signal processing tasks. This embodiment of the present invention may acquire task input features corresponding to different speech signal processing tasks respectively for the sample speech, and different speech signal processing tasks may correspond to different task input features.

[0031] Step S110: The electronic device determines a target training loss function based on a training loss function of each speech signal processing task.

[0032] This embodiment of the present invention updates the parameters of the shared layer and each task layer of the multi-task neural network by training all the speech signal processing tasks. Therefore, a total training loss function (referred to as the target training loss function) used for training needs to be determined based on the training loss function of each speech signal processing task.

[0033] Optionally, in view of the conventional solution of training a neural network respectively for each speech signal processing task, this embodiment of the present invention may determine the training loss function of each

speech signal processing task. Therefore, for any speech signal processing task, this embodiment of the present invention may multiply the training loss function of the speech signal processing task by a weight corresponding to the speech signal processing task to obtain a corresponding multiplication result of the speech signal processing task, so as to determine a corresponding multiplication result of each speech signal processing task; and then obtain the target training loss function by adding the corresponding multiplication results.

[0034] As an example, assuming that a training loss function of an i^{th} speech signal processing task is L_i , and a_i is a weight corresponding to the i^{th} speech signal processing task, the target training loss function L_{all} may be determined according to the following formula:

$$L_{all} = \sum_{i=1}^N \alpha_i L_i$$

where the value of a_i may be set according to an actual situation, or may be set to 1; and N is a total number of speech signal processing tasks.

[0035] Step S120: The electronic device uses the task input feature of each speech signal processing task of the sample speech as a training input of the multi-task neural network, and updates parameters of the shared layer and each task layer of the multi-task neural network by minimizing the target training loss function as a training objective, until the multi-task neural network converges, to obtain a speech signal processing model.

[0036] After determining the task input feature of each speech signal processing task of the sample speech and determining the target training loss function for training, this embodiment of the present invention may train the multi-task neural network to update the parameters of the shared layer and each task layer of the multi-task neural network. Specifically, this embodiment of the present invention may use the task input feature of each speech signal processing task of the sample speech as a training input of the multi-task neural network, and train the multi-task neural network by minimizing the target training loss function as a training objective to update the parameters of the shared layer and each task layer of the multi-task neural network, until the multi-task neural network converges, to obtain the speech signal processing model. When the multi-task neural network satisfies a convergence condition, the multi-task neural network converges. The convergence condition may include but is not limited to: the number of iterations of the training reaches a maximum number, or the target training loss function no longer decreases, and the like. This is not specifically limited in this embodiment of the present invention.

[0037] Optionally, after determining the training input and the training objective, this embodiment of the present invention may update the parameters of the shared layer

and each task layer of the multi-task neural network by using a stochastic gradient descent (SGD) method and/or a back propagation (BP) method.

[0038] Optionally, in the process of updating the parameters of the shared layer and each task layer of the multi-task neural network by minimizing the target training loss function as the training objective, the parameters of the shared layer may be updated based on the target training loss function. For example, during each training, the stochastic gradient descent method may be used to update the parameters of the shared layer based on the target training loss function obtained through the each training. The parameters of the task layer corresponding to any speech signal processing task may be updated based on the loss function of the speech signal processing task. For example, during each training, the stochastic gradient descent method may be used to update the parameters of the task layer corresponding to the speech signal processing task based on the training loss function of the speech signal processing task obtained through the each training. Therefore, not only the shared layer can reflect the association between the speech signal processing tasks having common features, but also each task layer can reflect the task feature of the corresponding speech signal processing task, so that an output result of each task layer can better reflect a task requirement of the corresponding speech signal processing task.

[0039] Optionally, in an example, the shared layer may be an LSTM network, and a task layer may be a fully connected MLP network. For example, updating the parameters of the shared layer of the multi-task neural network may be updating parameters of the LSTM network, including but not limited to updating, in the LSTM network, connection parameters from an input layer to a hidden layer, connection parameters from the hidden layer to an output layer or connection parameters between the hidden layers. For example, updating the parameters of a task layer of the multi-task neural network may be updating parameters of the fully connected MLP network, including but not limited to updating, in the fully connected MLP network, connection parameters from an input layer to a hidden layer or connection parameters from the hidden layer to an output layer.

[0040] Optionally, for ease of understanding, in an optional example, if the weight corresponding to each speech signal processing task is set to 1, and the plurality of speech signal processing tasks includes the echo cancellation task and the speech detection task, the process of training the speech signal processing model may include, as shown in FIG. 6.

using input features of the echo cancellation task and the speech detection task of the sample speech as a training input of the multi-task neural network; and updating parameters of the shared layer, the echo cancellation task layer and the speech detection task layer of the multi-task neural network by minimizing a sum of the training loss function of the echo cancellation task and the training loss function of the speech detection task as a training

objective, until the number of iterations of the multi-task neural network reaches a maximum number or the sum of the training loss function of the echo cancellation task and the training loss function of the speech detection task no longer decreases, to obtain the speech signal processing model.

[0041] Specifically, during each training, the parameters of the shared layer of the multi-task neural network may be updated based on the sum of the training loss function of the echo cancellation task and the training loss function of the speech detection task that are obtained through the each training; the parameters of the echo cancellation task layer may be updated based on the training loss function of the echo cancellation task obtained through the each training; and the parameters of the speech detection task layer may be updated based on the training loss function of the speech detection task obtained through the each training.

[0042] Optionally, generally the training loss function of the echo cancellation task may be, for example, a difference between estimated clean-speech spectral energy and an actual value; and the training loss function of the speech detection task may be, for example, a difference between an estimated speech presence probability and an actual value. Correspondingly, if the weight corresponding to each speech signal processing task is set to 1, the target training loss function may be determined as the sum of the training loss function of the echo cancellation task and the training loss function of the speech detection task. Therefore, during the training of the multi-task neural network, minimizing of the sum of the training loss function of the echo cancellation task and the training loss function of the speech detection task may be used as the training objective. The minimizing of the sum of the training loss function of the echo cancellation task and the training loss function of the speech detection task may specifically be minimizing a sum of the difference between the estimated clean-speech spectral energy and the actual value and the difference between the estimated speech presence probability and the actual value.

[0043] As can be seen, in the speech signal processing model training method shown in FIG. 5, the parameters of the shared layer and each task layer of the multi-task neural network may be updated based on the multi-task neural network including the shared layer and the task layer corresponding to each speech signal processing task by using the task input feature of each speech signal processing task of the sample speech as the training input, to obtain the speech signal processing model by training. Because this embodiment of the present invention simultaneously updates, based on the multi-task neural network including the shared layer and the task layer corresponding to each speech signal processing task, the parameters of the shared layer and each task layer of the multi-task neural network by using the task input feature of each speech signal processing task of the sample speech, instead of training a neural network

respectively for each speech signal processing task, the computational complexity of training the speech signal processing model is greatly and effectively reduced, and the efficiency of training the speech signal processing model is improved.

[0044] Compared with the conventional manner of training a neural network respectively for each speech signal processing task, the foregoing manner of training the multi-task neural network based on all the speech signal processing tasks simultaneously to update the parameters of the shared layer and each task layer of the multi-task neural network can reduce the computational complexity. Further, this embodiment of the present invention further provides a solution of training the multi-task neural network by stages. This solution is obtained based on the task feature of each speech signal processing task in the speech signal processing procedure, and can avoid large differences between speech signal processing tasks in the speech signal processing procedure. In addition, this solution may train the multi-task neural network by using some speech signal processing tasks, thereby ensuring the convergence of the parameters of the multi-task neural network.

[0045] Optionally, FIG. 7 shows another optional procedure of a speech signal processing model training method according to an embodiment of the present invention. The method may be applied to an electronic device having a data processing capability. Referring to FIG. 7, the procedure may include the following steps.

[0046] Step S200: The electronic device acquires a sample speech.

[0047] Step S210: The electronic device determines at least one first-class speech signal processing task from a plurality of speech signal processing tasks of a speech signal processing procedure.

[0048] Optionally, in an implementation, the first-class speech signal processing task may be a basic task in the plurality of speech signal processing tasks involved in the speech signal processing procedure. It may be understood that the basic task may be considered as a task providing an auxiliary effect for other speech signal processing tasks in the plurality of speech signal processing tasks of the speech signal processing procedure.

[0049] In an optional example, the plurality of speech signal processing tasks includes an echo cancellation task and a speech detection task. The echo cancellation task can estimate a single-channel speech spectrum and greatly improve the accuracy of speech probability estimation. Therefore, the echo cancellation task may be considered as a basic speech signal processing task.

[0050] Optionally, in another optional implementation, the first-class speech signal processing task may be considered as a task having a higher training complexity in the plurality of speech signal processing tasks involved in the speech signal processing procedure.

[0051] A process of determining the first-class speech signal processing task may be as follows: when the training complexity of the speech signal processing task is

higher than a set complexity threshold, determining that the speech signal processing task is the first-class speech signal processing task; otherwise, determining that the speech signal processing task is not the first-class speech signal processing task.

[0052] In an optional example, the plurality of speech signal processing tasks includes an echo cancellation task and a speech detection task. Clean-speech energy values of all M frequency bands are required for the single-channel speech spectrum estimation performed by the echo cancellation task, where M is generally a positive integer greater than 1. For example, the value of M may be 512. A single value estimate indicating whether a current frame includes a speech is required for the speech presence probability estimation performed by the speech detection task, where M is much greater than 1. Because the training complexity of the echo cancellation task is much higher than that of the speech detection task, the echo cancellation task may be considered as a first-class speech signal processing task having a higher training complexity.

[0053] In this embodiment of the present invention, the quantity of first-class speech signal processing tasks may be one or more.

[0054] Step S220: The electronic device determines a task input feature of the first-class speech signal processing task of the sample speech and a task input feature of each speech signal processing task of the sample speech.

[0055] After determining the first-class speech signal processing task, this embodiment of the present invention may determine, for the sample speech, the task input feature of the first-class speech signal processing task of the sample speech. The determining the task input feature of the first-class speech signal processing task may be determining a task input feature of the echo cancellation task of the sample speech. In addition, for each speech signal processing task involved in the speech signal processing procedure, this embodiment of the present invention determines the task input feature of each speech signal processing task of the sample speech. The determining the task input feature of each speech signal processing task may be determining the task input feature of the echo cancellation task of the sample speech, a task input feature of the speech detection task, and the like.

[0056] Step S230: The electronic device determines a first target training loss function based on a training loss function of the first-class speech signal processing task; and determines a target training loss function based on a training loss function of each speech signal processing task.

[0057] Optionally, for the first-class speech signal processing task, this embodiment of the present invention may determine the training loss function of the first-class speech signal processing task. In a case that the quantity of first-class speech signal processing tasks is at least one, for any first-class speech signal processing

task, this embodiment of the present invention may multiply the training loss function of the first-class speech signal processing task by a weight corresponding to the first-class speech signal processing task, to obtain a multiplication result of the first-class speech signal processing task, so as to determine the multiplication result of each first-class speech signal processing task; and then obtain the first target training loss function by adding the multiplication result of each first-class speech signal processing task.

[0058] Optionally, assuming that a training loss function of an i^{th} first-class speech signal processing task is $L1_i$, and $\alpha 1_i$ is a weight corresponding to the i^{th} first-class speech signal processing task, the first target training loss function $L1_{all}$ may be determined according to the following formula:

$$L1_{all} = \sum_{i=1}^{N1} \alpha 1_i L1_i,$$

where N1 is a total number of first-class speech signal processing tasks.

[0059] Optionally, this embodiment of the present invention may process each speech signal processing task to obtain the target training loss function with reference to the foregoing step S110.

[0060] Step S240: The electronic device uses the task input feature of the first-class speech signal processing task of the sample speech as a training input of a multi-task neural network, and updates parameters of a shared layer and a task layer corresponding to the first-class speech signal processing task of the multi-task neural network by minimizing the first target training loss function as a training objective, until the multi-task neural network converges, to obtain a first multi-task neural network.

[0061] Optionally, this embodiment of the present invention may first train an initial multi-task neural network based on the task input feature of the first-class speech signal processing task by minimizing the first target training loss function as the training objective.

[0062] During specific training, the parameters of the shared layer and the task layer corresponding to the first-class speech signal processing task of the multi-task neural network may be updated. A specific process of updating the parameters may be as follows: the electronic device uses the task input feature of the first-class speech signal processing task of the sample speech as a training input of the multi-task neural network, and updates the parameters of the shared layer and the task layer corresponding to the first-class speech signal processing task of the multi-task neural network through a plurality of iterations by minimizing the first target training loss function as a training objective, until the number of iterations reaches a maximum number or the first target training loss function no longer decreases, so as to obtain the first multi-task neural network.

[0063] Optionally, this embodiment of the present invention may update the parameters of the shared layer based on the first target training loss function obtained through each training. For the task layer corresponding to each first-class speech signal processing task, the parameters of the task layer corresponding to the first-class speech signal processing task may be updated based on the training loss function of the first-class speech signal processing task obtained through each training.

[0064] Step S250: The electronic device uses the task input feature of each speech signal processing task of the sample speech as a training input of the first multi-task neural network, and updates parameters of the shared layer and a task layer corresponding to each speech signal processing task of the multi-task neural network by minimizing the target training loss function as a training objective, until the first multi-task neural network converges, to obtain a speech signal processing model.

[0065] Optionally, this embodiment of the present invention trains the multi-task neural network based on the task input feature of the first-class speech signal processing task by minimizing the first target training loss function as the training objective, to obtain the first multi-task neural network, and then trains the first multi-task neural network based on the task input feature of each speech signal processing task by minimizing the target training loss function as the training objective, to obtain the speech signal processing model.

[0066] During specific training, the parameters of the shared layer and the task layer corresponding to each speech signal processing task of the first multi-task neural network may be updated. A specific process of updating the parameters may be: using the task input feature of each speech signal processing task of the sample speech as the training input of the first multi-task neural network, and updating the parameters of the shared layer and the task layer corresponding to each speech signal processing task of the first multi-task neural network through iterations by minimizing the target training loss function as the training objective, until the number of iterations reaches a maximum number, or the target training loss function no longer decreases, so as to obtain the speech signal processing model.

[0067] Optionally, this embodiment of the present invention may update the parameters of the shared layer based on the target training loss function obtained through each training. For the task layer corresponding to each speech signal processing task, the parameters of the task layer corresponding to the speech signal processing task may be updated based on the training loss function of the speech signal processing task obtained through each training.

[0068] For ease of understanding the sequential training processes shown in step S240 and step S250, in an optional example, if the weight corresponding to each speech signal processing task is set to 1, and the plurality of speech signal processing tasks includes the echo can-

cellation task and the speech detection task, the process of training the speech signal processing model according to this embodiment of the present invention may include, as shown in FIG. 8:

first using an input feature of the echo cancellation task of the sample speech as a training input of the multi-task neural network; and updating parameters of the shared layer and the task layer corresponding to the echo cancellation task of the multi-task neural network by minimizing of a training loss function of the echo cancellation task as a training objective, until the number of iterations of the multi-task neural network reaches a maximum number or the training loss function of the echo cancellation task no longer decreases, to obtain the first multi-task neural network, where the input feature of the echo cancellation task may be spectral energy of a noisy single-channel speech of the sample speech and spectral energy of a speech marked as clean, and the training objective may be minimizing a difference between estimated clean-speech spectral energy and an actual value; and

then, using input features of the echo cancellation task and the speech detection task of the sample speech as a training input of the first multi-task neural network; and updating parameters of the shared layer, the echo cancellation task layer and the speech detection task layer of the first multi-task neural network by minimizing a sum of the training loss function of the echo cancellation task and the training loss function of the speech detection task as a training objective, until the number of iterations of the first multi-task neural network reaches a maximum number or the sum of the training loss function of the echo cancellation task and the training loss function of the speech detection task no longer decreases, to obtain the speech signal processing model.

[0069] As can be seen, this embodiment of the present invention may determine, from the plurality of speech signal processing tasks, the basic task or a task having a higher training complexity based on the speech signal processing model training method shown in FIG. 7, to obtain at least one first-class speech signal processing task. Therefore, this embodiment of the present invention first uses the task input feature of the first-class speech signal processing task as the training input of the multi-task neural network to perform an updating training on the parameters of the shared layer and the task layer corresponding to the first-class speech signal processing task of the multi-task neural network, to obtain the first multi-task neural network; and then uses the task input feature of each speech signal processing task as the training input of the first multi-task neural network to perform an updating training on the parameters of the shared layer and each task layer of the first multi-task neural

network, to obtain the speech signal processing model by training.

[0070] In this process, because the neural network training is not performed on each speech signal processing task respectively, the computational complexity of training the speech signal processing model is reduced. In addition, training the multi-task neural network by first using the input feature of the first-class speech signal processing task and then using the task input feature of each speech signal processing task as the training input enables the training process to reflect the association between the speech signal processing tasks, and ensures the effective convergence of the parameters of the multi-task neural network and the reliable performance of the trained speech signal processing model.

[0071] During the training process of the speech signal processing model, the method shown in FIG. 7 first updates the parameters of the shared layer and the task layer corresponding to the first-class speech signal processing task of the multi-task neural network based on the task input feature of the first-class speech signal processing task, to obtain the first multi-task neural network by training. In the training process of obtaining the first multi-task neural network, because the first-class speech signal processing task is the basic task or a task having a higher training complexity in the speech signal processing procedure, the effective convergence of the parameters of the task layer corresponding to the first-class speech signal processing task is particularly critical for the performance of the speech signal processing model obtained by subsequent training.

[0072] Optionally, this embodiment the present invention may further perform, in a plurality of stages, the convergence training of the parameters of the task layer corresponding to the first-class speech signal processing task based on different input features of the first-class speech signal processing task, to further ensure the effective convergence of the parameters of the task layer corresponding to the first-class speech signal processing task. Optionally, FIG. 9 shows yet another optional procedure of a speech signal processing model training method according to this embodiment of the present invention. The procedure shown in FIG. 9 is only optional. During the training of the first multi-task neural network, the first multi-task neural network may be directly trained based on all task input features of the first-class speech signal processing task without being trained in a plurality of stages as shown in FIG. 9.

[0073] Optionally, the method shown in FIG. 9 may be applied to an electronic device having a data processing capability. Referring to FIG. 9, the procedure may include the following steps.

[0074] Step S300: The electronic device acquires a sample speech.

[0075] Step S310: The electronic device determines at least one first-class speech signal processing task from a plurality of speech signal processing tasks of a speech signal processing procedure.

[0076] Optionally, the implementation process of step S310 is similar to the process of step S210. For details of step S310, reference can be made to the description of step S210, and the details are not described herein again.

[0077] Step S320: The electronic device determines a task input feature of the first-class speech signal processing task of the sample speech and a task input feature of each speech signal processing task of the sample speech, where the task input feature of the first-class speech signal processing task includes: a plurality of groups of task input features, and the quantity of features included in one group of task input features is at least one.

[0078] Optionally, in this embodiment of the present invention, each first-class speech signal processing task may have a plurality of groups of task input features, and the quantity of features included in each group of task input features may be at least one.

[0079] In an optional example, the first-class speech signal processing task includes an echo cancellation task. In this embodiment of the present invention, a plurality of groups of task input features may be set for the echo cancellation task. For example, a first group of task input features of the echo cancellation task includes spectral energy of a noisy single-channel speech and spectral energy of a speech marked as clean; a second group of task input features of the echo cancellation task includes spectral energy of a multi-channel speech; and a third group of task input features of the echo cancellation task includes the spectral energy of the multi-channel speech, spectral energy of a reference signal (such as music played by an intelligent speaker), and the like.

[0080] Step S330: The electronic device determines a first target training loss function based on a training loss function of the first-class speech signal processing task; and determines a target training loss function based on a training loss function of each speech signal processing task.

[0081] Optionally, the implementation process of step S330 is similar to the process of step S230. For details of step S330, reference can be made to the description of step S230, and the details are not described herein again.

[0082] Step S340: The electronic device selects a current group of task input features corresponding to a current training stage from the plurality of groups of task input features of the first-class speech signal processing task of the sample speech based on the current training stage; uses the current group of task input features as a training input of a multi-task neural network trained in a previous training stage; and updates parameters of a shared layer and a task layer corresponding to the first-class speech signal processing task of the multi-task neural network trained in the previous training stage by minimizing the first target training loss function as a training objective, until a multi-task neural network trained based on the last group of task input features reaches convergence, to obtain a first multi-task neural network.

[0083] Optionally, step S340 may be considered as an optional implementation in which the electronic device trains the multi-task neural network progressively in a plurality of training stages based on the plurality of groups of task input features of the first-class speech signal processing task of the sample speech, to obtain the first multi-task neural network. One training stage uses one group of task input features as a training input, and uses minimizing of the first target training loss function as the training objective. A process of training the multi-task neural network progressively in the plurality of training stages may be updating the parameters of the shared layer and the task layer corresponding to the first-class speech signal processing task of the multi-task neural network progressively in the plurality of training stages. In addition to being implemented by step S340, this embodiment of the present invention does not exclude other manners of training the multi-task neural network progressively in a plurality of training stages by using the plurality of groups of task input features of the first-class speech signal processing task of the sample speech.

[0084] Optionally, in step S340, this embodiment of the present invention may train the first multi-task neural network in the plurality of training stages, so as to train the multi-task neural network by respectively using each group of task input features of the first-class speech signal processing task as a training input according to the training stage, to obtain the first multi-task neural network. In addition, in the current training stage, the current group of task input features selected currently by the first-class speech signal processing task is used as the training input of the multi-task neural network trained in the previous training stage.

[0085] Optionally, in an example, the first-class speech signal processing task has three groups of task input feature, that is, a first group of task input features, a second group of task input features, and a third group of task input features. In this case, this embodiment of the present invention may first use the first group of task input features as a training input of a to-be-trained multi-task neural network, and updates the parameters of the shared layer and the task layer corresponding to the first-class speech signal processing task of the multi-task neural network by minimizing the first target training loss function as a training objective, until a multi-task neural network trained based on the first group of task input features reaches convergence, to obtain a multi-task neural network trained in a first training stage. A process of using the first group of task input features as the training input of the to-be-trained multi-task neural network may be as follows: for the first training stage, the selected task input feature of the current training stage is the first group of task input features.

[0086] Then, this embodiment of the present invention uses the second group of task input features as a training input of the multi-task neural network trained in the first training stage, and updates parameters of the shared layer and the task layer corresponding to the first-class

speech signal processing task of the multi-task neural network trained in the first training stage by minimizing the first target training loss function as a training objective, until a multi-task neural network trained based on the second group of task input features reaches convergence, to obtain a multi-task neural network trained in the second training stage. A process of using the second group of task input features as the training input of the multi-task neural network trained in the first training stage may be as follows: for the second training stage, the selected task input feature of the current training stage is the second group of task input features.

[0087] Then, this embodiment of the present invention uses the third group of task input features as a training input of the multi-task neural network trained in the second training stage, and updates parameters of the shared layer and the task layer corresponding to the first-class speech signal processing task of the multi-task neural network trained in the second training stage by minimizing the first target training loss function as a training objective, until a multi-task neural network trained based on the third group of task input features reaches convergence, to obtain the first multi-task neural network, thereby completing the process of training the first multi-task neural network in a plurality of training stages based on the plurality of groups of task input features of the first-class speech signal processing task. A process of using the third group of task input features as the training input of the multi-task neural network trained in the second training stage may be as follows: for a third training stage, the selected task input feature of the current training stage is the third group of task input features.

[0088] For ease of understanding, assuming that the first-class speech signal processing task is an echo cancellation task, the first group of task input features of the echo cancellation task includes spectral energy of a noisy single-channel speech and spectral energy of a speech marked as clean; the second group of task input features of the echo cancellation task includes spectral energy of a multi-channel speech; and the third group of task input features of the echo cancellation task includes the spectral energy of the multi-channel speech, spectral energy of a reference signal and the like. The spectral energy of the reference signal may be music played by an intelligent speaker.

[0089] Correspondingly, this embodiment of the present invention may first use the spectral energy of the noisy single-channel speech and the spectral energy of the speech marked as clean of the sample speech as a training input of the multi-task neural network, and updates the parameters of the shared layer and the task layer of the echo cancellation task of the multi-task neural network by minimizing a difference between estimated clean-speech spectral energy and an actual value as a training objective, until the number of iterations reaches a maximum number or the training objective no longer decreases.

[0090] Then, this embodiment of the present invention

uses the spectral energy of the multi-channel speech of the sample speech as a training input of the multi-task neural network trained in the previous stage, and updates the parameters of the shared layer and the task layer of the echo cancellation task of the multi-task neural network by minimizing the difference between the estimated clean-speech spectral energy and the actual value as a training objective, until the number of iterations reaches a maximum number or the training objective no longer decreases, so that the trained multi-task neural network has a multi-channel spatial filtering capability.

[0091] After completing the multi-channel training, this embodiment of the present invention may further use the spectral energy of the multi-channel speech and the spectral energy of the reference signal of the sample speech as a training input of the multi-task neural network trained in the previous stage, and updates the parameters of the shared layer and the task layer of the echo cancellation task of the multi-task neural network by minimizing the difference between the estimated clean-speech spectral energy and the actual value as a training objective, until the number of iterations reaches a maximum number or the training objective no longer decreases, to obtain the first multi-task neural network, so that the first multi-task neural network can better fit a multi-channel input signal and the reference signal.

[0092] Optionally, the foregoing example of the plurality of groups of task input features of the first-class speech signal processing task is only optional. In this embodiment of the present invention, the quantity of groups of task input features of the first-class speech signal processing task and specific features included in each group of task input features may be set according to a specific situation. For example, in the foregoing example, the task input features including the spectral energy of the noisy single-channel speech, the spectral energy of the speech marked as clean and the spectral energy of the multi-channel speech may also be combined for training.

[0093] Step S350: The electronic device uses the task input feature of each speech signal processing task of the sample speech as a training input of the first multi-task neural network, and updates parameters of the shared layer and a task layer corresponding to each speech signal processing task of the first multi-task neural network by minimizing the target training loss function as a training objective, until the first multi-task neural network converges, to obtain a speech signal processing model.

[0094] Optionally, the implementation process of step S350 is similar to the process of step S250. For details of step S350, reference can be made to the description of step S250, and the details are not described herein again.

[0095] Optionally, after the first multi-task neural network is obtained, speech signal processing tasks such as speech detection, direction detection and de-reverberation are simple tasks independent of each other, and

may be combined for training. Therefore, after the first multi-task neural network is obtained, the task input features of each speech signal processing task of the sample speech may be combined for training the first multi-task neural network, to obtain the speech signal processing model.

[0096] For all the foregoing training manners, in the process of updating the parameters of the shared layer and a task layer, the parameters of the shared layer are updated based on a sum of training loss functions of all tasks used for the current training; and the parameters of the task layer are updated based on a training loss function of a task corresponding to the task layer. Therefore, the trained speech signal processing model not only can reflect an association between the speech signal processing tasks having common features through the shared layer, but also can reflect the task feature of the corresponding speech signal processing task through each task layer.

[0097] To summarize the training processes of the speech signal processing model described above, a basic core procedure of the speech signal processing model training method according to the embodiments of the present invention may be as shown in FIG. 10. FIG. 10 shows still another optional procedure of a speech signal processing model training method according to an embodiment of the present invention. Referring to FIG. 10, the procedure may include the following steps.

[0098] Step S400: An electronic device acquires a speech sample and determines a task input feature of each speech signal processing task of the speech sample.

[0099] Optionally, for details of step S400, reference can be made to the description of step S100. The implementation process of step S400 is similar to the process of step S100, and the details are not described herein again.

[0100] Step S410: The electronic device determines a target training loss function based on a training loss function of each speech signal processing task.

[0101] Optionally, for details of step S410, reference can be made to the description of step S110. The implementation process of step S410 is similar to the process of step S110, and the details are not described herein again.

[0102] Step S420: The electronic device uses the task input feature of each speech signal processing task of the sample speech as a training input of a to-be-trained multi-task neural network, and updates parameters of a shared layer and each task layer of the to-be-trained multi-task neural network by minimizing the target training loss function as a training objective, until the to-be-trained multi-task neural network reaches convergence, to obtain a speech signal processing model.

[0103] Optionally, in an optional implementation, in step S420, the to-be-trained multi-task neural network may be an initial multi-task neural network (the corresponding process may be implemented by the procedure

shown in FIG. 5).

[0104] Optionally, in another implementation, in step S420, the to-be-trained multi-task neural network may be a first multi-task neural network. This embodiment of the present invention may first obtain the first multi-task neural network through training using the method procedure of training the first multi-task neural network as shown in FIG. 7, and use the first multi-task neural network as the to-be-trained multi-task neural network; then use the task input feature of each speech signal processing task of the sample speech as a training input of the first multi-task neural network by using the method shown in FIG. 10; and update parameters of the shared layer and each task layer of the first multi-task neural network by minimizing the target training loss function as a training objective, until the first multi-task neural network converges, to obtain the speech signal processing model.

[0105] Optionally, the first multi-task neural network may be trained based on the task input feature of the first-class speech signal processing task of the sample speech. Further, in an optional example, the first-class speech signal processing task may have a plurality of groups of task input features. This embodiment of the present invention may train the first multi-task neural network in a plurality of training stages based on the training procedure of the first multi-task neural network as shown in FIG. 9.

[0106] Regardless of whether the foregoing to-be-trained multi-task neural network is the initial multi-task neural network or the first multi-task neural network, the structure of the to-be-trained multi-task neural network includes a shared layer and a task layer corresponding to each speech signal processing task. For the shared layer, the parameters of the shared layer are updated based on the target training loss function by minimizing the target training loss function as the training objective. For the task layer corresponding to any speech signal processing task, the parameters of the task layer of the speech signal processing task are updated based on the training loss function of the speech signal processing task by minimizing the target training loss function as the training objective.

[0107] The speech signal processing model training method according to this embodiment of the present invention may obtain the speech signal processing model through training based on the multi-task neural network including the shared layer and the task layer corresponding to each speech signal processing task, instead of training a neural network respectively for each speech signal processing task, thereby effectively reducing the computational complexity of training the speech signal processing model, and improving the efficiency of training.

[0108] Further, in the training process, the speech signal processing model is first trained based on the task input feature of the first-class speech signal processing task of the sample speech, and then trained based on the task input feature of each speech signal processing

task. Therefore, an association between a plurality of tasks in the speech signal processing procedure can be found, thereby improving the performance of speech signal processing, and ensuring the reliable performance of the trained speech signal processing model.

[0109] After obtaining the speech signal processing model through training by using the foregoing method, optionally, the embodiments of the present invention may use the speech signal processing model to replace the conventional speech signal processing procedure of the terminal. For example, the output result of each task layer of the speech signal processing model may be specifically used to replace the conventional task processing result of the speech signal processing task corresponding to each task layer of the terminal.

[0110] In another implementation, the embodiments of the present invention may use the speech signal processing model to assist the conventional speech signal processing procedure of the terminal. For example, the output of each task layer of the speech signal processing model may be specifically used to assist task processing in the conventional corresponding speech signal processing task of the terminal.

[0111] FIG. 11 is a schematic diagram of an application scenario of a speech signal processing model. As shown in FIG. 11, after the speech signal processing model is obtained through training, in the embodiments of the present invention, speech signal processing of a front end may be performed on a to-be-recognized speech input to an instant messaging client by using the speech signal processing model, and then the processed speech is transmitted to a speech backend server of an instant messaging application for speech recognition. Optionally, the instant messaging client may respectively use the output of each task layer of the speech signal processing model for the to-be-recognized speech as an assisting processing signal of the corresponding speech signal processing task, to assist processing of each speech signal processing task, thereby improving the output accuracy of the result of each speech signal processing task.

[0112] Referring to FIG. 11, as an optional application scenario, on a basis that the instant messaging client loads the trained speech signal processing model of the embodiments of the present invention, a specific application process may include the following steps.

[0113] S1. An instant messaging client acquires a to-be-recognized speech input.

[0114] S2. The instant messaging client determines, based on a to-be-trained speech signal processing model, an output result of each task layer of the speech signal processing model for the to-be-recognized speech.

[0115] The speech signal processing model is obtained through training a multi-task neural network by minimizing a target training loss function as a training objective. The target training loss function is determined based on a training loss function of each speech signal processing task. The multi-task neural network includes a shared layer and a task layer corresponding to each

speech signal processing task.

[0116] S3. The instant messaging client uses the output result of each task layer for the to-be-recognized speech as a task processing result of the speech signal processing task corresponding to each task layer, or uses the output result of each task layer for the to-be-recognized speech to assist task processing in the corresponding speech signal processing task, so as to obtain a speech signal processing result of a front end.

[0117] S4. The instant messaging client sends the speech signal processing result of the front end to a speech backend server, so that the speech backend server performs speech recognition on the to-be-recognized speech based on the speech signal processing result.

[0118] Optionally, FIG. 12 is a schematic diagram of using an output result of a speech signal processing model. Referring to FIG. 12, for a conventional echo cancellation task of the terminal, an echo cancellation task layer of the speech signal processing model may output a speech spectrum estimation result of the to-be-recognized speech, so as to use the speech spectrum estimation result as an assisting processing signal of the conventional echo cancellation task of the terminal, so that the echo cancellation task may better distinguish a ratio of the reference signal and the speech signal during processing, thereby improving an accuracy of an output result of the echo cancellation task. Certainly, in another implementation, the embodiments of the present invention may also directly use the output result of the echo cancellation task layer of the speech signal processing model for the to-be-recognized speech as the output result of the echo cancellation task.

[0119] For a speech detection task, a speech detection task layer of the speech signal processing model may output an output result of the to-be-recognized speech, and use the output result as an assisting processing signal of the conventional speech detection task of the terminal, so as to improve an accuracy of the output result of the speech detection task. A weighted average value of an output result of the speech detection task layer and the output result of the conventional speech detection task of the terminal may be used as a final output result of the speech detection task. Certainly, in another implementation, the embodiments of the present invention may also directly use the output result of the speech detection task layer of the speech signal processing model for the to-be-recognized speech as the output result of the speech detection task.

[0120] For a speech direction detection task, a speech direction detection task layer of the speech signal processing model may output an output result of the to-be-recognized speech, so that the output result is used to assist the conventional speech direction detection task of the terminal to estimate the speech and noise of the to-be-recognized speech, so as to obtain a more accurate speech direction estimation result. The output result of the to-be-recognized speech outputted by the speech

direction detection task layer may be: the speech/noise spectrum estimation result of the to-be-recognized speech. Certainly, in another implementation, the embodiments of the present invention may also directly use the speech/noise spectrum estimation result of the to-be-recognized speech outputted by the speech direction detection task layer as the output result of speech direction detection task.

[0121] For a microphone array enhancement task, a microphone array enhancement task layer of the speech signal processing model may output a speech/noise spectrum of the to-be-recognized speech, to assist the conventional microphone array enhancement task of the terminal to more accurately estimate a target direction of an array algorithm and parameters such as a noise covariance matrix needed in the array algorithm. Certainly, in another implementation, the embodiments of the present invention may also directly use the output result of the microphone array enhancement task layer as the output result of the microphone array enhancement task.

[0122] For a single-channel noise reduction task, a single-channel noise reduction task layer of the speech signal processing model may output the speech/noise spectrum of the to-be-recognized speech, to assist the conventional single-channel noise reduction task of the terminal to obtain key parameters needed in the single-channel noise reduction task such as a signal-noise ratio, to improve the processing effect of the single-channel noise reduction task. Certainly, in another implementation, the embodiments of the present invention may also directly use the output result of the single-channel noise reduction task layer as the output result of the single-channel noise reduction task.

[0123] For a de-reverberation task, a de-reverberation task layer of the speech signal processing model may output a room reverberation estimation, to assist the conventional de-reverberation task of the terminal to regulate the parameters of the algorithm, so as to control a level of de-reverberation. Certainly, in another implementation, the embodiments of the present invention may also directly use the output result of the de-reverberation task layer as the output result of the de-reverberation task.

[0124] Optionally, the application of the speech signal processing model to the speech signal processing procedure of the to-be-recognized speech is merely an example, which may be understood as an application of the speech signal processing procedure in an intelligent speaker scenario. Certainly, in different application scenarios, an application manner of the speech signal processing model may be adjusted according to an actual situation, but do not separate from a thought that uses the speech signal processing model to replace the conventional speech signal processing procedure, or, uses the speech signal processing model to assist the conventional speech signal processing procedure.

[0125] The following describes a speech signal processing model training apparatus provided in the embodiments of the present invention. The speech signal

processing model training apparatus described below may be considered as, a program module needed to be arranged by an electronic device to implement the speech signal processing model training method provided in the embodiments of the present invention. The speech signal processing model training apparatus described below may correspond to the speech signal processing model training method described above.

[0126] FIG. 13 is a structural block diagram of a speech signal processing model training apparatus according to an embodiment of the present invention. The apparatus may be applied to an electronic device having a data processing capability. Referring to FIG. 13, the apparatus may include:

a task input feature determining module 100, configured to acquire a sample speech and determine a task input feature of each speech signal processing task of the sample speech;

a target loss function determining module 200, configured to determine a target training loss function based on a training loss function of each speech signal processing task; and

a model training module 300, configured to use the task input feature of each speech signal processing task of the sample speech as a training input of a to-be-trained multi-task neural network, and update parameters of a shared layer and each task layer of the to-be-trained multi-task neural network by minimizing the target training loss function as a training objective, until the to-be-trained multi-task neural network converges, to obtain a speech signal processing model,

the to-be-trained multi-task neural network including: the shared layer and the task layer corresponding to each speech signal processing task.

[0127] Optionally, the model training module 300 being configured to update parameters of a shared layer and each task layer of the to-be-trained multi-task neural network by minimizing the target training loss function as a training objective specifically includes:

for the shared layer, updating the parameters of the shared layer based on the target training loss function by minimizing the target training loss function as the training objective; and for the task layer corresponding to any speech signal processing task, updating the parameters of the task layer of the speech signal processing task based on the training loss function of the speech signal processing task by minimizing the target training loss function as the training objective.

[0128] Optionally, the to-be-trained multi-task neural network may include a first multi-task neural network. Correspondingly, FIG. 14 is another structural block diagram of a speech signal processing model training ap-

paratus according to an embodiment of the present invention. With reference to FIG. 13 and FIG. 14, the apparatus may further include:

a first network training module 400, configured to determine at least one first-class speech signal processing task from a plurality of speech signal processing tasks of a speech signal processing procedure; determine a task input feature of the first-class speech signal processing task of the sample speech; determine a first target training loss function based on a training loss function of the first-class speech signal processing task; and use the task input feature of the first-class speech signal processing task of the sample speech as a training input of an initial multi-task neural network, and update parameters of a shared layer and a task layer corresponding to the first-class speech signal processing task of the initial multi-task neural network by minimizing the first target training loss function as a training objective, until the initial multi-task neural network converges, to obtain the first multi-task neural network.

[0129] Optionally, the first network training module 400 being configured to determine at least one first-class speech signal processing task from a plurality of speech signal processing tasks of a speech signal processing procedure specifically includes:

determining a basic task from the plurality of speech signal processing tasks, and determining the basic task as the first-class speech signal processing task, the basic task being a task providing an auxiliary effect for other speech signal processing tasks in the plurality of speech signal processing tasks.

[0130] Optionally, the first network training module 400 being configured to determine at least one first-class speech signal processing task from a plurality of speech signal processing tasks of a speech signal processing procedure specifically includes:

determining, from the plurality of speech signal processing tasks, a speech signal processing task having a training complexity higher than a set complexity threshold as the first-class speech signal processing task.

[0131] Optionally, the first network training module 400 being configured to determine a first target training loss function based on a training loss function of the first-class speech signal processing task specifically includes:

for any first-class speech signal processing task, multiplying the training loss function of the first-class speech signal processing task by a weight corresponding to the first-class speech signal processing task to obtain a multiplication result of the first-class speech signal processing task, so as to determine a multiplication result of each first-class speech signal processing task; and

obtaining the first target training loss function by adding the multiplication result of each first-class speech signal processing task.

[0132] Optionally, the task input feature of the first-class speech signal processing task of the sample speech includes: a plurality of groups task input features. The quantity of features included in one group of task input features is at least one.

[0133] Correspondingly, the first network training module 400 being configured to use the task input feature of the first-class speech signal processing task of the sample speech as a training input of an initial multi-task neural network, and update parameters of a shared layer and a task layer corresponding to the first-class speech signal processing task of the initial multi-task neural network by minimizing the first target training loss function as a training objective, until the initial multi-task neural network converges, to obtain the first multi-task neural network specifically includes:

training the initial multi-task neural network progressively based on the plurality of groups of task input features of the first-class speech signal processing task of the sample speech in a plurality of training stages, to obtain the first multi-task neural network, one training stage using one group of task input features of the first-class speech signal processing task of the sample speech as a training input, and minimizing the first target training loss function as a training objective.

[0134] Optionally, the first network training module 400 being configured to train the initial multi-task neural network progressively based on the plurality of groups of task input features of the first-class speech signal processing task of the sample speech in a plurality of training stages, to obtain the first multi-task neural network specifically includes:

selecting a current group of task input features corresponding to a current training stage from the plurality of groups of task input features of the first-class speech signal processing task of the sample speech based on the current training stage; and using the current group of task input features as a training input of the multi-task neural network trained in a previous training stage, and updating parameters of the shared layer and the task layer corresponding to the first-class speech signal processing task of the multi-task neural network trained in the previous training stage by minimizing the first target training loss function as a training objective, until a multi-task neural network trained based on a last group of task input features reaches convergence, to obtain the first multi-task neural network.

[0135] Optionally, in another implementation, the to-be-trained multi-task neural network may include an initial multi-task neural network. The target loss function determining module 200 being configured to determine a target training loss function based on a training loss function of each speech signal processing task specifically includes:

for any speech signal processing task, multiplying the training loss function of the speech signal processing task by a weight corresponding to the

speech signal processing task to obtain a corresponding multiplication result of the speech signal processing task, so as to determine a corresponding multiplication result of each speech signal processing task; and

obtaining the target training loss function by adding the corresponding multiplication result of each speech signal processing task.

[0136] Optionally, the shared layer in the multi-task neural network may include an LSTM network. Each task layer of the multi-task neural network may include a fully connected MLP network corresponding to each speech signal processing task.

[0137] Optionally, the model training module 300 being configured to update parameters of a shared layer and each task layer of the to-be-trained multi-task neural network may specifically include:

updating, in the LSTM network of the to-be-trained multi-task neural network, connection parameters from an input layer to a hidden layer, connection parameters from the hidden layer to an output layer or connection parameters between the hidden layers; and updating, in the fully connected MLP network corresponding to each speech signal processing task, connection parameters from an input layer to a hidden layer or connection parameters from the hidden layer to an output layer.

[0138] Optionally, after training the speech signal processing model, the model training module 300 may be applied in the speech signal processing procedure at a speech front end. Optionally, FIG. 15 is yet another structural block diagram of a speech signal processing model training apparatus according to an embodiment of the present invention. With reference to FIG. 14 and FIG. 15, the apparatus may further include:

a model application module 500, configured to determine an output result of each task layer of the speech signal processing model for the to-be-recognized speech; and use the output result of each task layer for the to-be-recognized speech as a task processing result of the speech signal processing task corresponding to each task layer.

[0139] Optionally, the model application module 500 is further configured to determine an output result of each task layer of the speech signal processing model for the to-be-recognized speech; and use the output result of each task layer for the to-be-recognized speech to assist task processing in the speech signal processing task corresponding to each task layer.

[0140] Optionally, the model application module 500 may also be used in the apparatus shown in FIG. 13.

[0141] The speech signal processing model training apparatus provided by the embodiments of the present invention may be applied to an electronic device. Optionally, a hardware structure of the electronic device may be shown in FIG. 16, and includes at least one processor 1, at least one communications interface 2, at least one

memory 3, and at least one communications bus 4.

[0142] In this embodiment of the present invention, the quantity of each of the processor 1, the communications interface 2, the memory 3, and the communications bus 4 is at least one, and communication among the processor 1, the communications interface 2, and the memory 3 is implemented by using the communications bus 4. Optionally, the processor 1 may be a central processing unit (CPU) or an application-specific integrated circuit (ASIC), or may be configured as one or more integrated circuits for implementing the embodiments of the present invention. The memory 3 may include a high-speed RAM memory, or may further include a non-volatile memory, for example, at least one magnetic disk memory.

[0143] The memory stores a program, the processor invokes the program stored by the memory, and the program is configured for:

acquiring a sample speech and determining a task input feature of each speech signal processing task of the sample speech; determining a target training loss function based on a training loss function of each speech signal processing task; and using the task input feature of each speech signal processing task of the sample speech as a training input of a to-be-trained multi-task neural network, and updating parameters of a shared layer and each task layer of the to-be-trained multi-task neural network by minimizing the target training loss function as a training objective, until the to-be-trained multi-task neural network converges, to obtain a speech signal processing model,

the to-be-trained multi-task neural network including: the shared layer and the task layer corresponding to each speech signal processing task.

[0144] Optionally, the program is further configured for:

for the shared layer, updating the parameters of the shared layer based on the target training loss function by minimizing the target training loss function as the training objective; and

for the task layer corresponding to any speech signal processing task, updating the parameters of the task layer of the speech signal processing task based on the training loss function of the speech signal processing task by minimizing the target training loss function as the training objective.

[0145] Optionally, the program is further configured for:

determining at least one first-class speech signal processing task from a plurality of speech signal processing tasks of a speech signal processing procedure;

determining a task input feature of the first-class speech signal processing task of the sample speech;

determining a first target training loss function based on a training loss function of the first-class speech signal processing task; and

using the task input feature of the first-class speech signal processing task of the sample speech as a training input of an initial multi-task neural network, and updating parameters of a shared layer and a task layer corresponding to the first-class speech signal processing task of the initial multi-task neural network by minimizing the first target training loss function as a training objective, until the initial multi-task neural network converges, to obtain the first multi-task neural network.

[0146] Optionally, the program is further configured for:

for any first-class speech signal processing task, multiplying the training loss function of the first-class speech signal processing task by a weight corresponding to the first-class speech signal processing task to obtain a multiplication result of the first-class speech signal processing task, so as to determine a multiplication result of each first-class speech signal processing task; and

obtaining the first target training loss function by adding the multiplication result of each first-class speech signal processing task.

[0147] Optionally, the program is further configured for: determining a basic task from the plurality of speech signal processing tasks, and determining the basic task as the first-class speech signal processing task, the basic task being a task providing an auxiliary effect for other speech signal processing tasks in the plurality of speech signal processing tasks.

[0148] Optionally, the program is further configured for: determining, from the plurality of speech signal processing tasks, a speech signal processing task having a training complexity higher than a set complexity threshold as the first-class speech signal processing task.

[0149] Optionally, the program is further configured for:

training the initial multi-task neural network progressively based on the plurality of groups of task input features of the first-class speech signal processing task of the sample speech in a plurality of training stages, to obtain the first multi-task neural network,

one training stage using one group of task input features of the first-class speech signal processing task of the sample speech as a training input, and minimizing the first target training loss function as a training objective.

[0150] Optionally, the program is further configured for:

selecting a current group of task input features corresponding to a current training stage from the plurality of groups of task input features of the first-class speech signal processing task of the sample speech based on the current training stage; and

using the current group of task input features as a training input of the multi-task neural network trained in a previous training stage, and updating parameters of the shared layer and the task layer corresponding to the first-class speech signal processing task of the multi-task neural network trained in the previous training stage by minimizing the first target training loss function as a training objective, until a multi-task neural network trained based on a last group of task input features reaches convergence, to obtain the first multi-task neural network.

[0151] Optionally, the program is further configured for:

for any speech signal processing task, multiplying the training loss function of the speech signal processing task by a weight corresponding to the speech signal processing task to obtain a corresponding multiplication result of the speech signal processing task, so as to determine a corresponding multiplication result of each speech signal processing task; and

obtaining the target training loss function by adding the corresponding multiplication result of each speech signal processing task.

[0152] Optionally, the program is further configured for:

updating, in the LSTM network of the to-be-trained multi-task neural network, connection parameters from an input layer to a hidden layer, connection parameters from the hidden layer to an output layer or connection parameters between the hidden layers; and

updating, in the fully connected MLP network corresponding to each speech signal processing task, connection parameters from an input layer to a hidden layer or connection parameters from the hidden layer to an output layer.

[0153] Optionally, the program is further configured for:

determining an output result of each task layer of the speech signal processing model for a to-be-recognized speech;

using the output result of each task layer for the to-be-recognized speech as a task processing result of

the speech signal processing task corresponding to each task layer.

[0154] Optionally, the program is further configured for:

determining an output result of each task layer of the speech signal processing model for a to-be-recognized speech;

using the output result of each task layer for the to-be-recognized speech to assist task processing in the speech signal processing task corresponding to each task layer.

[0155] The refinement function and extension function of the program may refer to the corresponding parts described above.

[0156] Further, the embodiments of the present invention further provide a storage medium, storing a program configured to be executed by a processor, the program being configured for:

acquiring a sample speech and determining a task input feature of each speech signal processing task of the sample speech; determining a target training loss function based on a training loss function of each speech signal processing task; and using the task input feature of each speech signal processing task of the sample speech as a training input of a to-be-trained multi-task neural network, and updating parameters of a shared layer and each task layer of the to-be-trained multi-task neural network by minimizing the target training loss function as a training objective, until the to-be-trained multi-task neural network converges, to obtain a speech signal processing model, the to-be-trained multi-task neural network including: the shared layer and the task layer corresponding to each speech signal processing task.

[0157] Optionally, the refinement function and extension function of the program may refer to the corresponding parts described above.

[0158] The embodiments in this specification are all described in a progressive manner. Description of each of the embodiments focuses on differences from other embodiments, and reference may be made to each other for the same or similar parts among respective embodiments. The apparatus embodiments are substantially similar to the method embodiments and therefore are only briefly described, and reference may be made to the method embodiments for the associated part.

[0159] Persons skilled in the art may further realize that, in combination with the embodiments herein, units and algorithm, steps of each example described can be implemented with electronic hardware, computer software, or the combination thereof. In order to clearly describe the interchangeability between the hardware and the software, compositions and steps of each example have been generally described according to functions in the foregoing descriptions. Whether the functions are executed in a mode of hardware or software depends on

particular applications and design constraint conditions of the technical solutions. Persons skilled in the art can use different methods to implement the described functions for each particular application, but it should not be considered that the implementation goes beyond the scope of the embodiments of the present disclosure.

[0160] In combination with the embodiments herein, steps of the method or algorithm described may be directly implemented using hardware, a software module executed by a processor, or the combination thereof. The software module may be placed in a random access memory (RAM), a memory, a read-only memory (ROM), an electrically programmable ROM (EPROM), an electrically erasable programmable ROM (EEPROM), a register, a hard disk, a removable magnetic disk, a CD-ROM, or any storage medium of other forms well-known in the technical field.

[0161] The above description of the disclosed embodiments enables persons skilled in the art to implement or use the present disclosure. Various modifications to these embodiments are obvious to persons skilled in the art, the general principles defined in the present disclosure may be implemented in other embodiments without departing from the core idea or scope of the present disclosure. Therefore, the present disclosure is not limited to these embodiments illustrated in the present disclosure, but needs to conform to the broadest scope consistent with the principles and novel features disclosed in the present disclosure.

Claims

1. A speech signal processing model training method, applied to an electronic device, the method comprising:

acquiring a sample speech and determining a task input feature of each speech signal processing task of the sample speech;
determining a target training loss function based on a training loss function of each speech signal processing task; and
using the task input feature of each speech signal processing task of the sample speech as a training input of a to-be-trained multi-task neural network, and updating parameters of a shared layer and each task layer of the to-be-trained multi-task neural network by minimizing the target training loss function as a training objective, until the to-be-trained multi-task neural network converges, to obtain a speech signal processing model,
the to-be-trained multi-task neural network comprising: the shared layer and the task layer corresponding to each speech signal processing task.

2. The speech signal processing model training method according to claim 1, wherein the updating parameters of a shared layer and each task layer of the to-be-trained multi-task neural network by minimizing the target training loss function as a training objective comprises:

for the shared layer, updating the parameters of the shared layer based on the target training loss function by minimizing the target training loss function as the training objective; and
for the task layer corresponding to any speech signal processing task, updating the parameters of the task layer of the speech signal processing task based on the training loss function of the speech signal processing task by minimizing the target training loss function as the training objective.

3. The speech signal processing model training method according to claim 1, wherein the to-be-trained multi-task neural network comprises a first multi-task neural network; and the method further comprises:

determining at least one first-class speech signal processing task from a plurality of speech signal processing tasks of a speech signal processing procedure;
determining a task input feature of the first-class speech signal processing task of the sample speech;
determining a first target training loss function based on a training loss function of the first-class speech signal processing task; and
using the task input feature of the first-class speech signal processing task of the sample speech as a training input of an initial multi-task neural network, and updating parameters of a shared layer and a task layer corresponding to the first-class speech signal processing task of the initial multi-task neural network by minimizing the first target training loss function as a training objective, until the initial multi-task neural network converges, to obtain the first multi-task neural network.

4. The speech signal processing model training method according to claim 3, wherein the determining at least one first-class speech signal processing task from a plurality of speech signal processing tasks of a speech signal processing procedure comprises: determining a basic task from the plurality of speech signal processing tasks, and determining the basic task as the first-class speech signal processing task, the basic task being a task providing an auxiliary effect for other speech signal processing tasks in the plurality of speech signal processing tasks.

5. The speech signal processing model training method according to claim 3, wherein the determining at least one first-class speech signal processing task from a plurality of speech signal processing tasks of a speech signal processing procedure comprises:
 - determining, from the plurality of speech signal processing tasks, a speech signal processing task having a training complexity higher than a set complexity threshold as the first-class speech signal processing task.
6. The speech signal processing model training method according to claim 1, wherein the to-be-trained multi-task neural network comprises an initial multi-task neural network; and the determining a target training loss function based on a training loss function of each speech signal processing task comprises:
 - for any speech signal processing task, multiplying the training loss function of the speech signal processing task by a weight corresponding to the speech signal processing task to obtain a corresponding multiplication result of the speech signal processing task, so as to determine a corresponding multiplication result of each speech signal processing task; and
 - obtaining the target training loss function by adding the corresponding multiplication result of each speech signal processing task.
7. The speech signal processing model training method according to claim 1, wherein the shared layer comprises a long short term memory (LSTM) network, and the task layer corresponding to each speech signal processing task comprises a fully connected multi layer perceptron (MLP) network corresponding to the speech signal processing task; and the updating parameters of a shared layer and each task layer of the to-be-trained multi-task neural network comprises:
 - updating, in the LSTM network of the to-be-trained multi-task neural network, connection parameters from an input layer to a hidden layer, connection parameters from the hidden layer to an output layer or connection parameters between the hidden layers; and
 - updating, in the fully connected MLP network corresponding to each speech signal processing task, connection parameters from an input layer to a hidden layer or connection parameters from the hidden layer to an output layer.
8. An electronic device, comprising: at least one memory and at least one processor; the memory storing a program, the processor invoking the program stored by the memory, and the program being con-

figured for:

acquiring a sample speech and determining a task input feature of each speech signal processing task of the sample speech; determining a target training loss function based on a training loss function of each speech signal processing task; and using the task input feature of each speech signal processing task of the sample speech as a training input of a to-be-trained multi-task neural network, and updating parameters of a shared layer and each task layer of the to-be-trained multi-task neural network by minimizing the target training loss function as a training objective, until the to-be-trained multi-task neural network converges, to obtain a speech signal processing model, the to-be-trained multi-task neural network comprising: the shared layer and the task layer corresponding to each speech signal processing task.

9. The electronic device according to claim 8, wherein the program is further configured for:
 - for the shared layer, updating the parameters of the shared layer based on the target training loss function by minimizing the target training loss function as the training objective; and
 - for the task layer corresponding to any speech signal processing task, updating the parameters of the task layer of the speech signal processing task based on the training loss function of the speech signal processing task by minimizing the target training loss function as the training objective.
10. The electronic device according to claim 8, wherein the program is further configured for:
 - determining at least one first-class speech signal processing task from a plurality of speech signal processing tasks of a speech signal processing procedure;
 - determining a task input feature of the first-class speech signal processing task of the sample speech;
 - determining a first target training loss function based on a training loss function of the first-class speech signal processing task; and
 - using the task input feature of the first-class speech signal processing task of the sample speech as a training input of an initial multi-task neural network, and updating parameters of a shared layer and a task layer corresponding to the first-class speech signal processing task of the initial multi-task neural network by minimizing the first target training loss function as a train-

ing objective, until the initial multi-task neural network converges, to obtain the first multi-task neural network.

11. The electronic device according to claim 10, wherein the program is further configured for:

for any first-class speech signal processing task, multiplying the training loss function of the first-class speech signal processing task by a weight corresponding to the first-class speech signal processing task to obtain a multiplication result of the first-class speech signal processing task, so as to determine a multiplication result of each first-class speech signal processing task; and obtaining the first target training loss function by adding the multiplication result of each first-class speech signal processing task.

12. The electronic device according to claim 10, wherein the program is further configured for:
determining a basic task from the plurality of speech signal processing tasks, and determining the basic task as the first-class speech signal processing task, the basic task being a task providing an auxiliary effect for other speech signal processing tasks in the plurality of speech signal processing tasks.

13. The electronic device according to claim 10, wherein the program is further configured for:
determining, from the plurality of speech signal processing tasks, a speech signal processing task having a training complexity higher than a set complexity threshold as the first-class speech signal processing task.

14. The electronic device according to claim 10, wherein the program is further configured for:

training the initial multi-task neural network progressively based on a plurality of groups of task input features of the first-class speech signal processing task of the sample speech in a plurality of training stages, to obtain the first multi-task neural network,
one training stage using one group of task input features of the first-class speech signal processing task of the sample speech as a training input, and minimizing the first target training loss function as a training objective.

15. The electronic device according to claim 12, wherein the program is further configured for:

selecting a current group of task input features corresponding to a current training stage from the plurality of groups of task input features of the first-class speech signal processing task of

the sample speech based on the current training stage; and

using the current group of task input features as a training input of the multi-task neural network trained in a previous training stage, and updating parameters of the shared layer and the task layer corresponding to the first-class speech signal processing task of the multi-task neural network trained in the previous training stage by minimizing the first target training loss function as a training objective, until a multi-task neural network trained based on a last group of task input features reaches convergence, to obtain the first multi-task neural network.

16. The electronic device according to claim 8, wherein the program is further configured for:

for any speech signal processing task, multiplying the training loss function of the speech signal processing task by a weight corresponding to the speech signal processing task to obtain a corresponding multiplication result of the speech signal processing task, so as to determine a corresponding multiplication result of each speech signal processing task; and obtaining the target training loss function by adding the corresponding multiplication result of each speech signal processing task.

17. The electronic device according to claim 8, wherein the program is further configured for:

updating, in the LSTM network of the to-be-trained multi-task neural network, connection parameters from an input layer to a hidden layer, connection parameters from the hidden layer to an output layer or connection parameters between the hidden layers; and
updating, in the fully connected MLP network corresponding to each speech signal processing task, connection parameters from an input layer to a hidden layer or connection parameters from the hidden layer to an output layer.

18. The electronic device according to claim 8, wherein the program is further configured for:

determining an output result of each task layer of the speech signal processing model for a to-be-recognized speech;
using the output result of each task layer for the to-be-recognized speech as a task processing result of the speech signal processing task corresponding to each task layer.

19. The electronic device according to claim 8, wherein the program is further configured for:

determining an output result of each task layer
of the speech signal processing model for a to-
be-recognized speech;
using the output result of each task layer for the
to-be-recognized speech to assist task process- 5
ing in the speech signal processing task corre-
sponding to each task layer.

20. A storage medium, storing a program configured to
be executed by a processor, the program being con- 10
figured for:

acquiring a sample speech and determining a
task input feature of each speech signal
processing task of the sample speech; 15
determining a target training loss function based
on a training loss function of each speech signal
processing task; and
using the task input feature of each speech sig-
nal processing task of the sample speech as a 20
training input of a to-be-trained multi-task neural
network, and updating parameters of a shared
layer and each task layer of the to-be-trained
multi-task neural network by minimizing the tar-
get training loss function as a training objective, 25
until the to-be-trained multi-task neural network
converges, to obtain a speech signal processing
model,
the to-be-trained multi-task neural network com-
prising: the shared layer and the task layer cor- 30
responding to each speech signal processing
task.

35

40

45

50

55

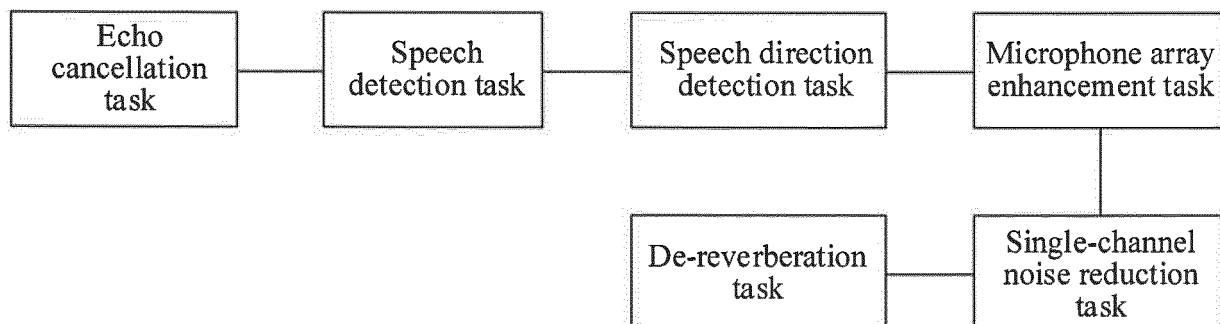


FIG. 1

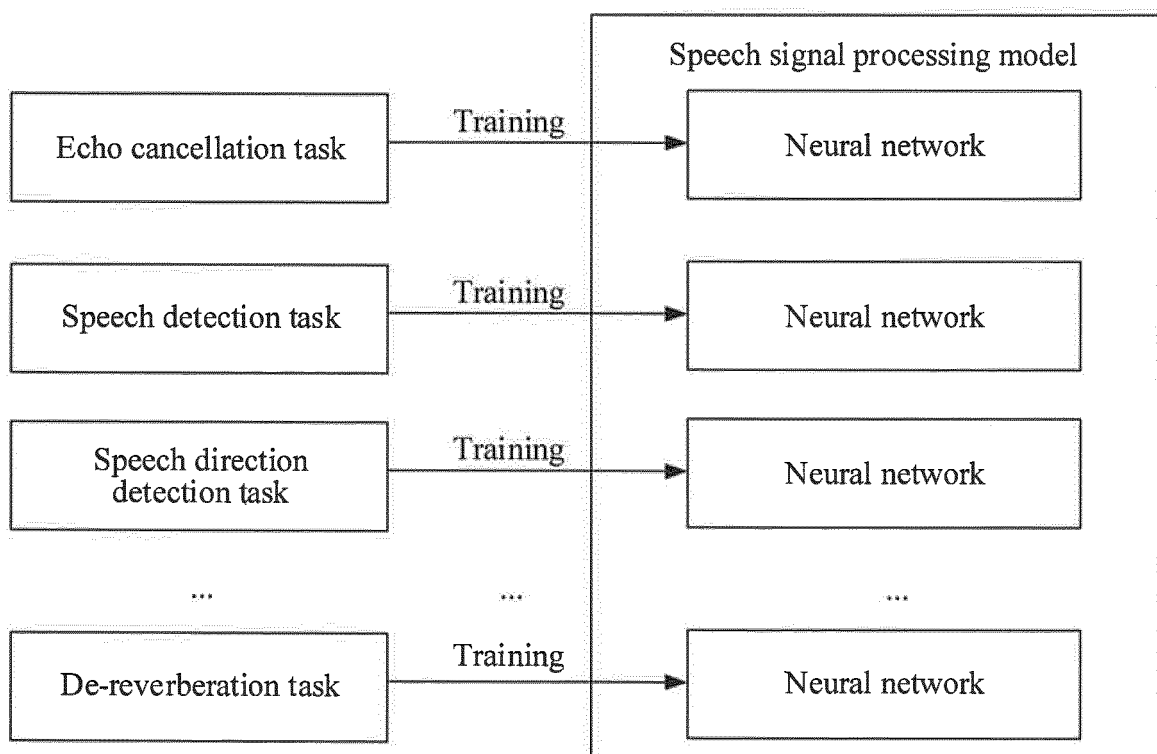


FIG. 2

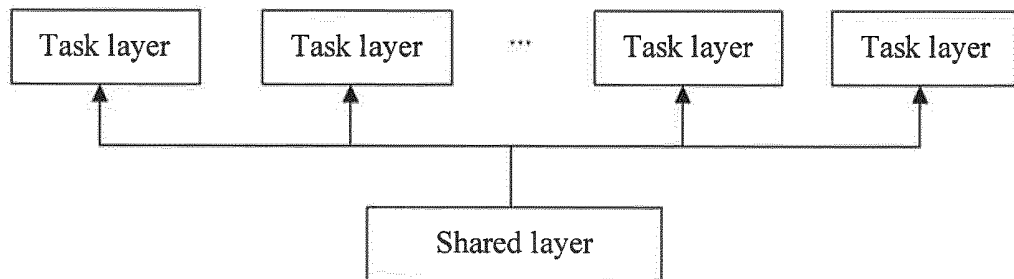


FIG. 3

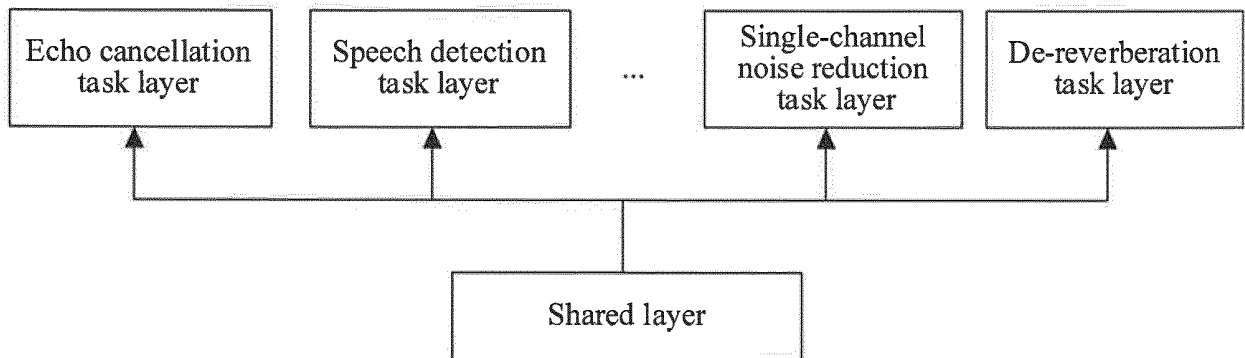


FIG. 4

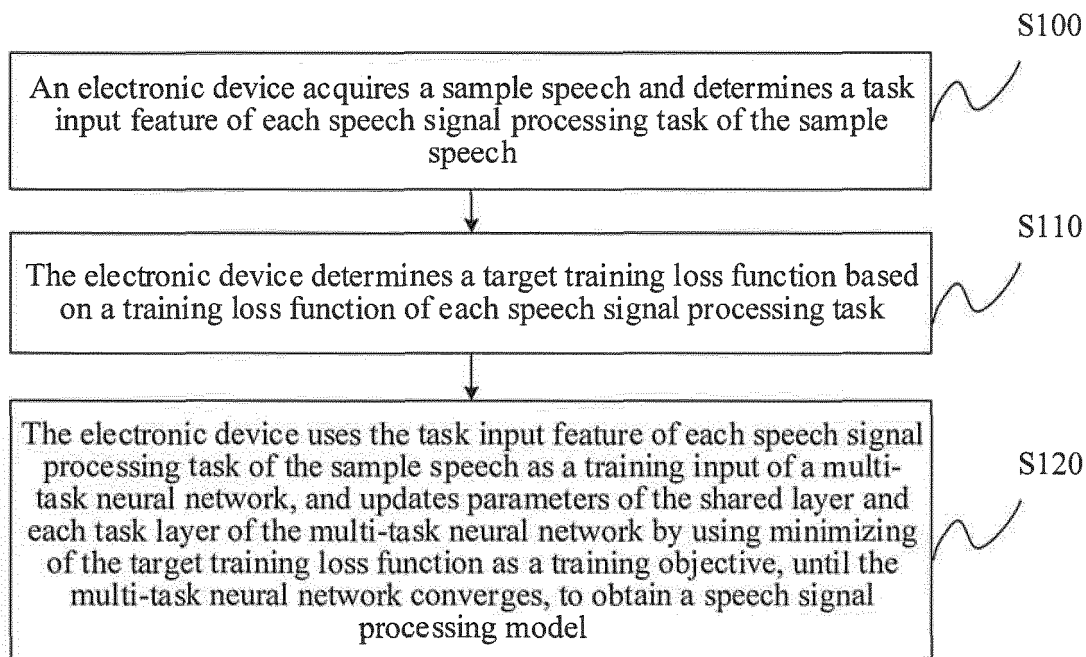


FIG. 5

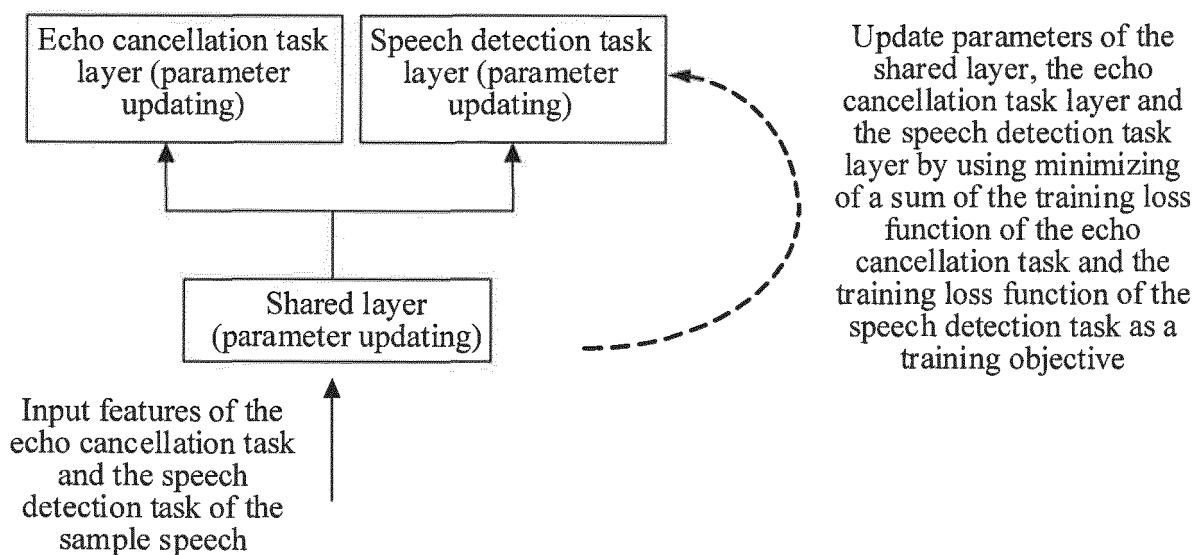


FIG. 6

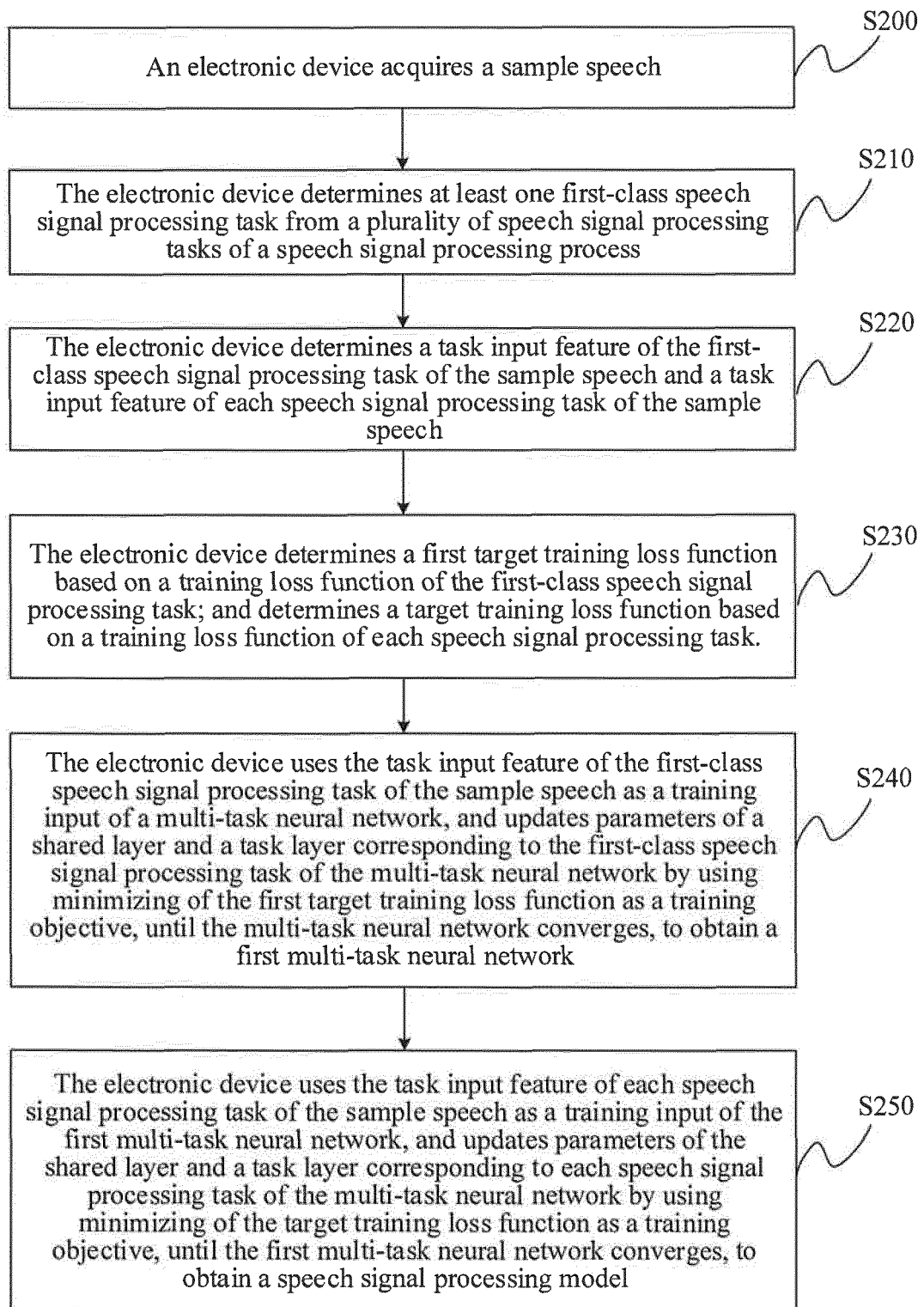


FIG. 7

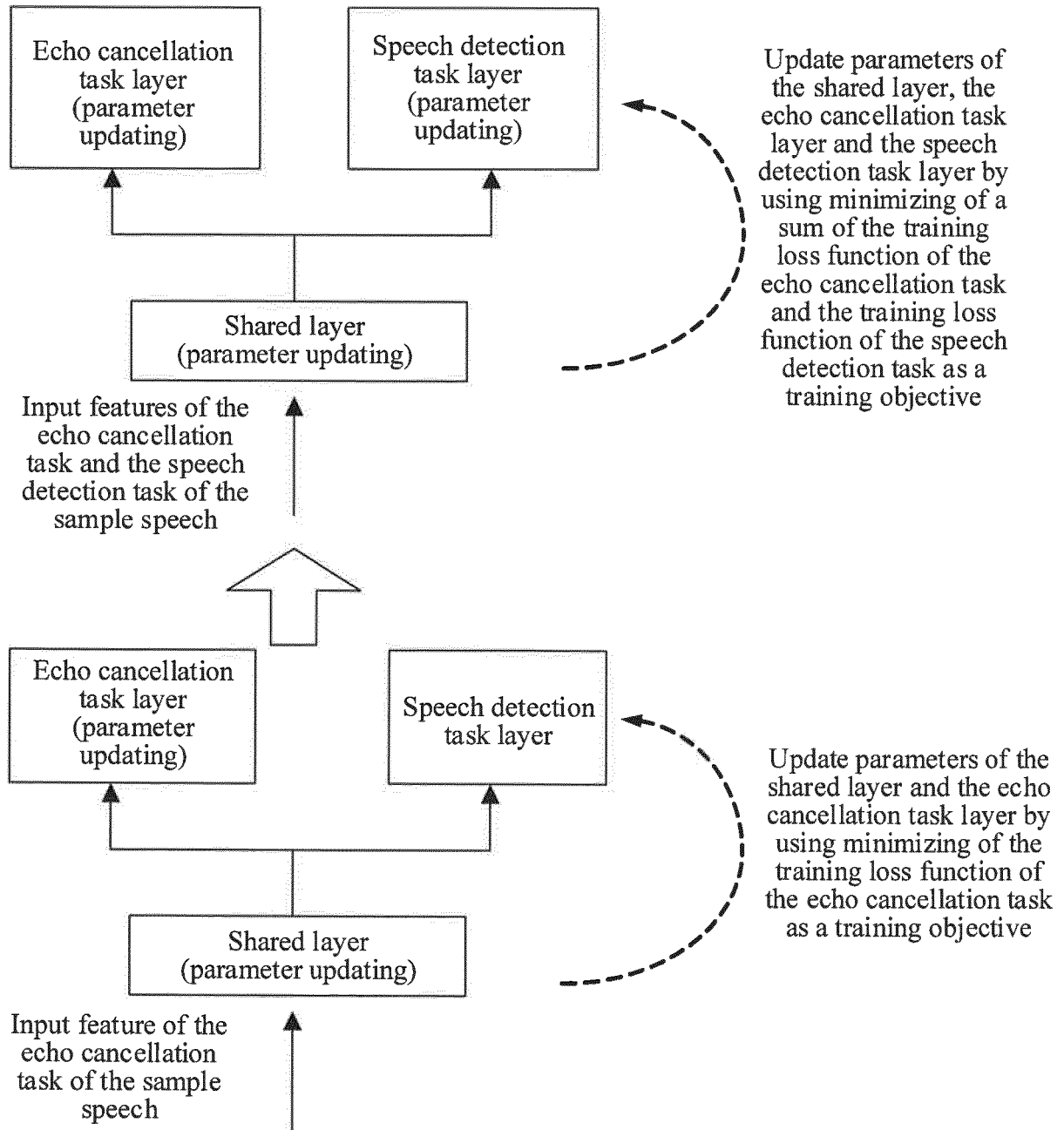


FIG. 8

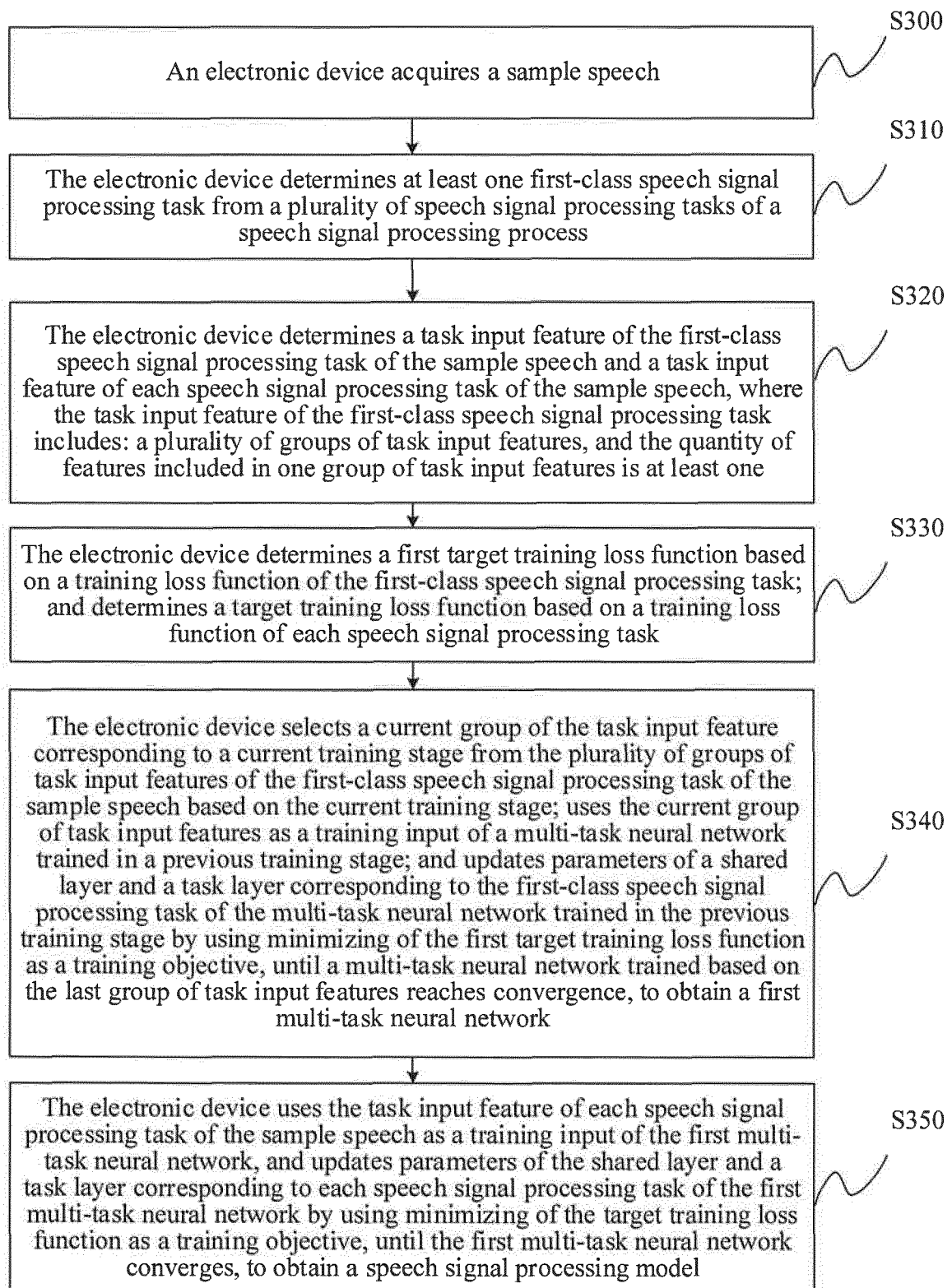


FIG. 9

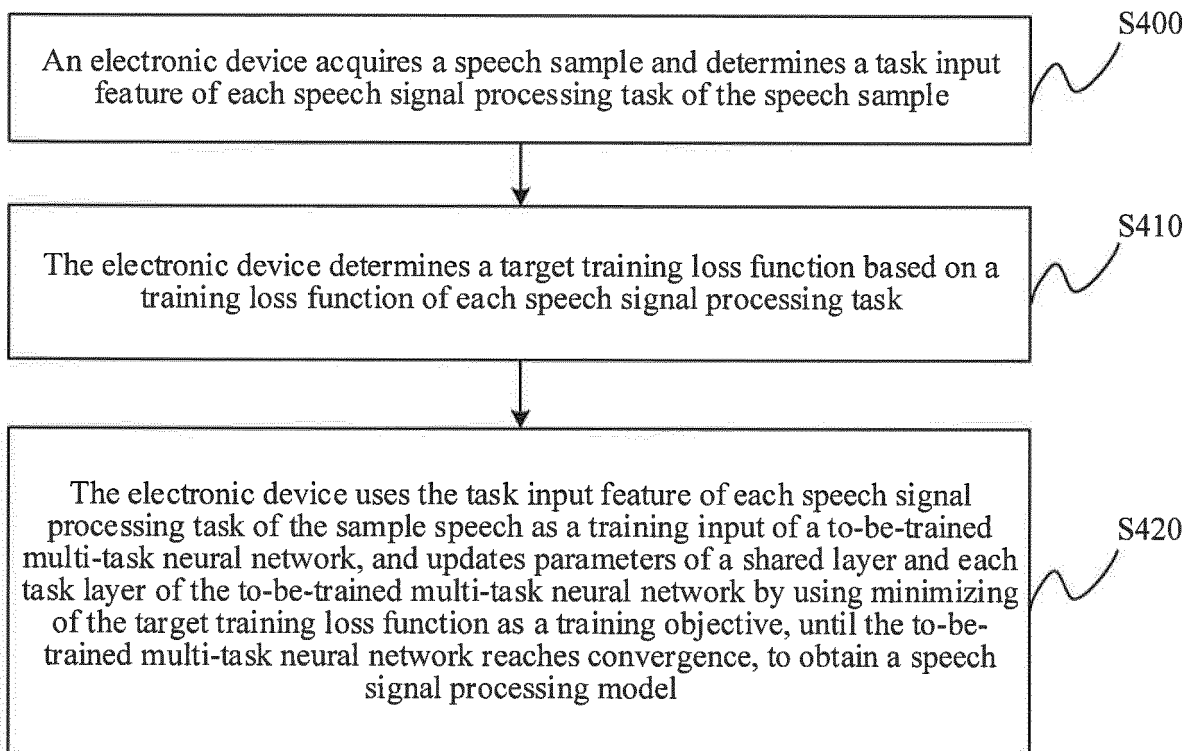


FIG. 10

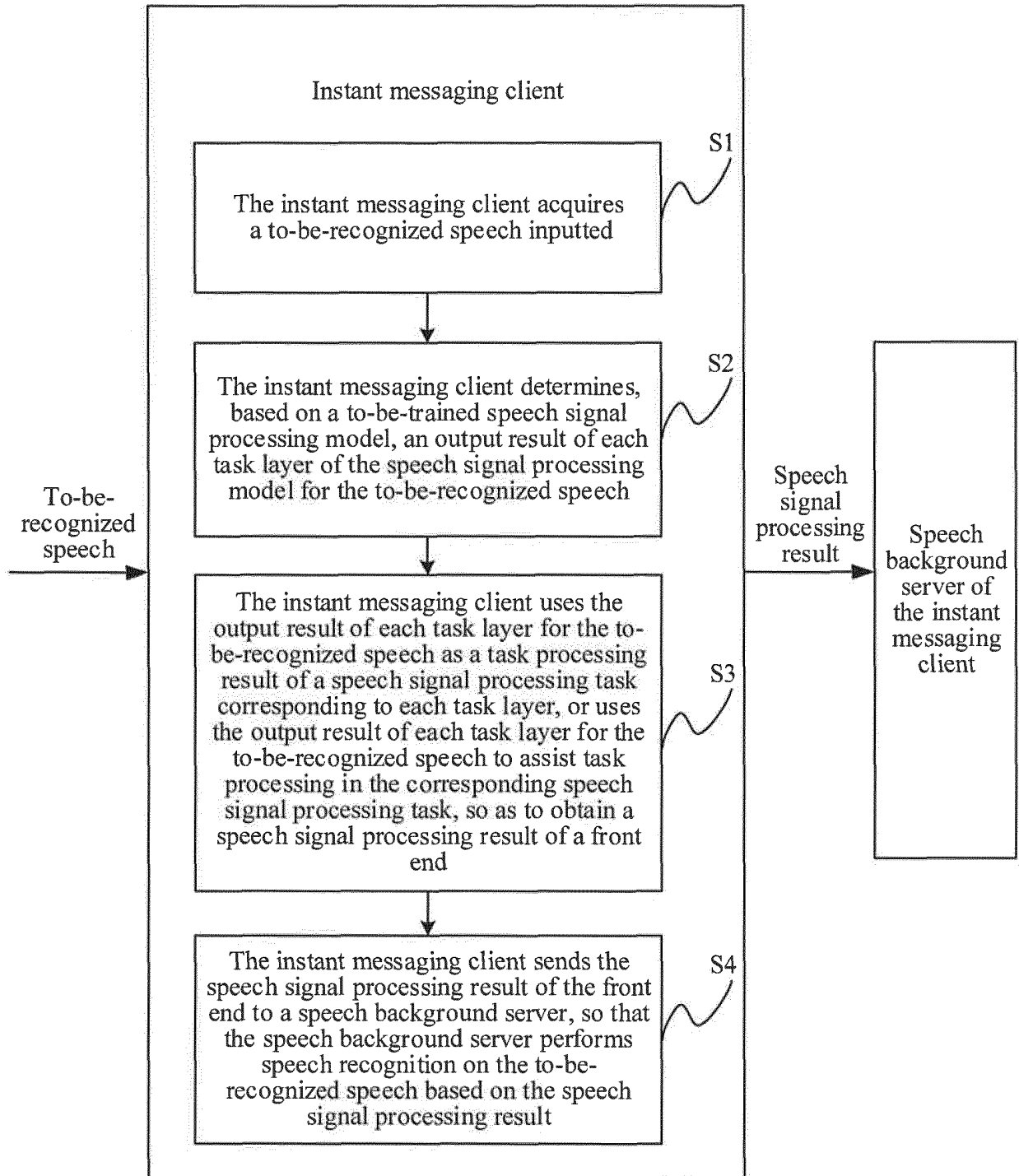


FIG. 11

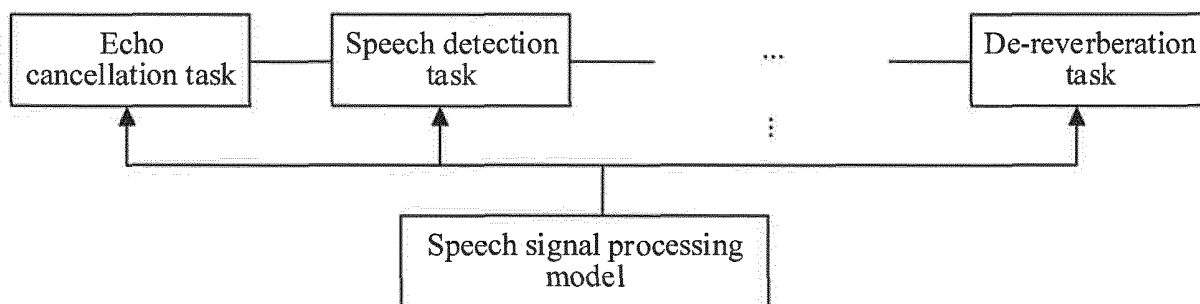


FIG. 12

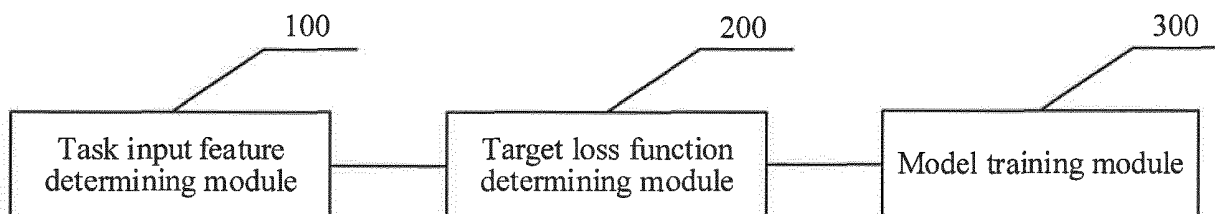


FIG. 13

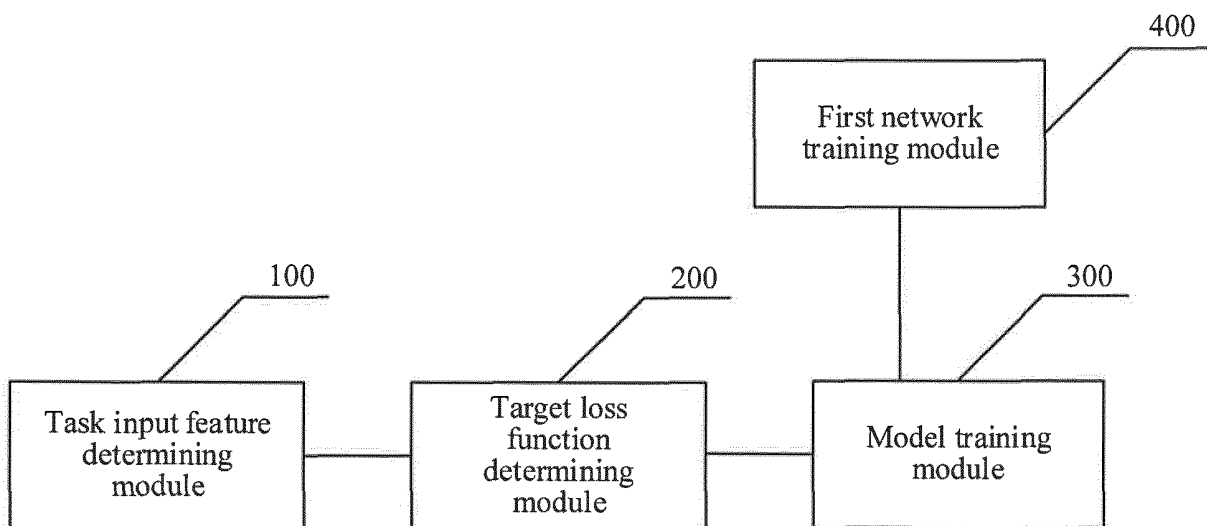


FIG. 14

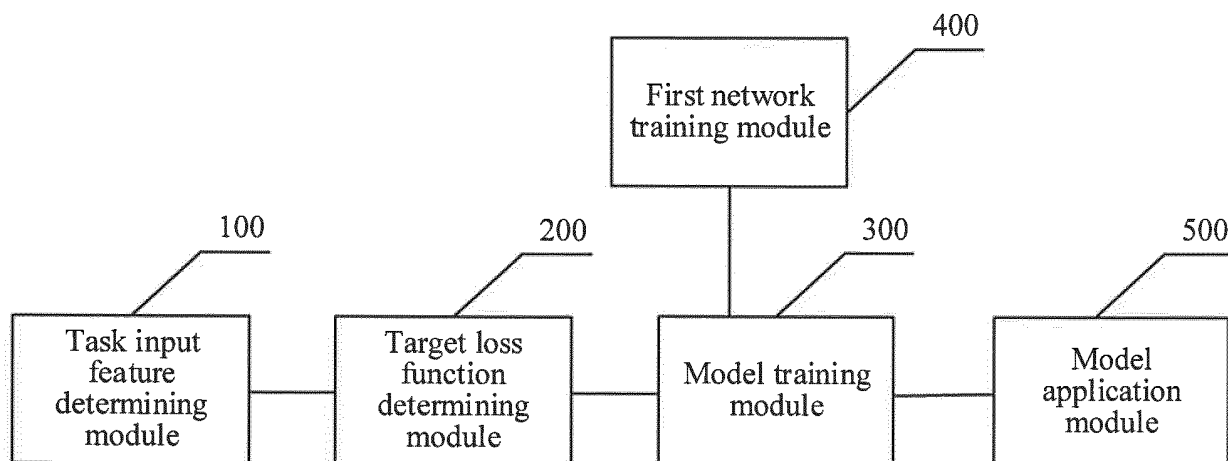


FIG. 15

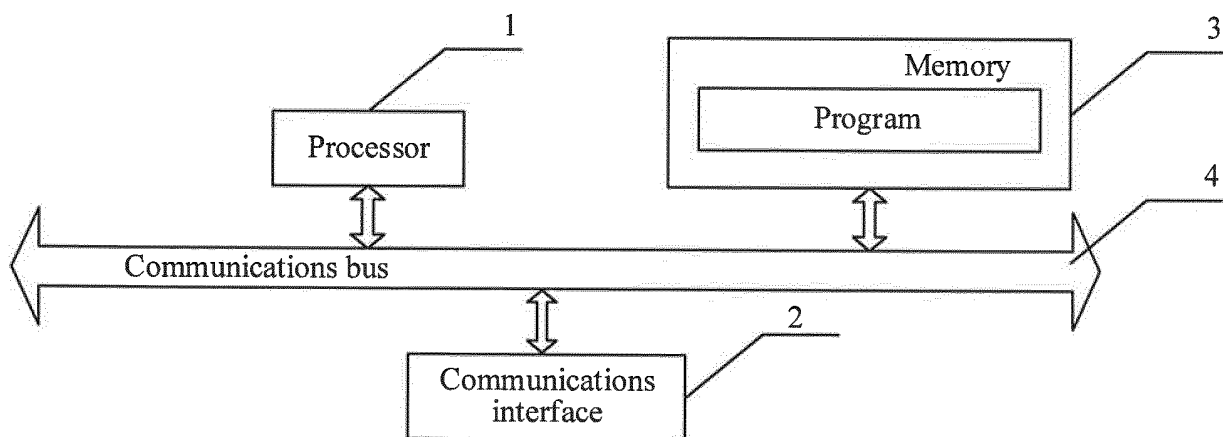


FIG. 16

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/115704

5	A. CLASSIFICATION OF SUBJECT MATTER G10L 17/04(2013.01)i; G10L 15/16(2006.01)i According to International Patent Classification (IPC) or to both national classification and IPC	
10	B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G10L Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched	
15	Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) CNPAT, CNKI, WPI, EPODOC: 语音, 声音, 音频, 训练, 长短时, 长短期, 记忆网络, 多层感知器, 反向传播, 深度, 神经网络, 卷积网络, 深度学习, 隐层, 回声消除, 语音检测, 降噪, 混响消除, 多任务神经网络, 损失函数, LSTM, Long Short Term Memory, MLP, Multi layer Perceptron, Back Propagation, BP, Stochastic Gradient Descent, SGD, speech, auditory, supervision label, time frequency, matrix, vector, DNN, lost, cost, multi-task, learning	
20	C. DOCUMENTS CONSIDERED TO BE RELEVANT	
	Category*	Citation of document, with indication, where appropriate, of the relevant passages Relevant to claim No.
25	A	CN 104881678 A (UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA) 02 September 2015 (2015-09-02) description, paragraphs [0012]-[0020] 1-20
30	A	CN 107357838 A (SHANGHAI JIAOTONG UNIVERSITY ET AL.) 17 November 2017 (2017-11-17) entire document 1-20
35	A	CN 106355248 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY) 25 January 2017 (2017-01-25) entire document 1-20
40	A	WO 2017083399 A2 (GOOGLE INC.) 18 May 2017 (2017-05-18) entire document 1-20
45	☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.	
50	* Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family
55	Date of the actual completion of the international search 18 December 2018	Date of mailing of the international search report 29 December 2018
	Name and mailing address of the ISA/CN State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088 China Facsimile No. (86-10)62019451	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/CN2018/115704

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	104881678	A	02 September 2015	None			
CN	107357838	A	17 November 2017	None			
CN	106355248	A	25 January 2017	None			
WO	2017083399	A2	18 May 2017	US	2017132513	A1	11 May 2017
				EP	3374933	A2	19 September 2018
				CN	108292374	A	17 July 2018

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- CN 201711191604 [0001]